Apprentissage sur corpus de relations lexicales sémantiques - La linguistique et l'apprentissage au service d'applications du traitement automatique des langues

Pascale Sébillot

HABILITATION À DIRIGER DES RECHERCHES

présentée devant

L'Université de Rennes 1
Institut de Formation Supérieure
en Informatique et en Communication

par

Pascale Sébillot

Apprentissage sur corpus de relations lexicales sémantiques
La linguistique et l'apprentissage au service d'applications du traitement automatique des langues

soutenue le 13 décembre 2002 devant le jury composé de

Mme	Laurence Danlos	Présidente et rapporteur
MM	Guy Lapalme	Rapporteur
	Jean Véronis	Rapporteur
	Pierre Zweigenbaum	Rapporteur
	Philippe Besnard	Examinateur
	Mohand Boughanem	Examinateur
	Olivier Ridoux	Examinateur

Remerciements

Je remercie très chaleureusement Laurence Danlos, Guy Lapalme, Jean Véronis et Pierre Zweigenbaum qui ont accepté la tâche de rapporter sur mes travaux. Laurence Danlos a également présidé le jury de ma soutenance et je la remercie aussi à ce titre.

Mohand Boughanem apporte à ce jury son expertise des domaines applicatifs que mes travaux abordent, et je le remercie beaucoup de sa présence.

J'adresse mes plus sincères remerciements à Philippe Besnard, pour sa participation à ce jury et pour l'intérêt qu'il a toujours porté à mes travaux, mais également pour sa présence amicale depuis mon arrivée à l'Irisa, en tant que responsable, membre ou collaborateur extérieur de mon équipe de recherche. Merci aussi à Olivier Ridoux, membre de ce jury, pour sa curiosité concernant tout ce qui a trait au traitement automatique des langues, qui nous a conduits à d'intéressantes discussions.

Cécile Fabre a relu les versions successives de mes chapitres, avec le recul et la rigueur que j'avais tant appréciés en elle lors de sa thèse. Je lui adresse mes remerciements les plus vifs, pour toutes ses suggestions et tout le temps qu'elle m'a consacré.

Merci également à Vincent Claveau pour la relecture attentive des parties de ce mémoire ayant trait au Lexique génératif, ainsi d'ailleurs que pour l'enthousiasme et le sérieux du travail réalisé depuis le début de sa thèse. Mathias Rossignol a joué un rôle équivalent pour les sections abordant la sémantique différentielle. Je lui adresse des remerciements tout aussi chaleureux.

Didier Bourigault et Ludovic Tanguy se sont plongés dans les écrits de F. Rastier pour répondre à mes questions : merci à eux ; et merci à Pierrette Bouillon pour sa disponibilité face à mes interrogations sur la théorie de J. Pustejovsky, mais également pour la collaboration sympathique que nous avons établie autour de ce formalisme.

Plusieurs personnes de l'Irisa m'ont fortement encouragée à rédiger ce document. Parmi elles, je voudrais citer et remercier Claude Labit et Daniel Herman qui, par leurs sollicitations amicales et répétées, ont su me conduire jusqu'à la réalisation de ce mémoire.

Je ne pourrais, et surtout ne voudrais, clore ces remerciements sans évoquer les membres de Repco, puis Aïda, et enfin du Patio (auxquels j'inclus bien évidemment les autres membres de TexMex, géographiquement dispersés dans les bâtiments de l'Irisa). Derrière ces appellations se trouvent des personnes qui ont su créer un environnement de travail chaleureux et stimulant que j'apprécie énormément. Ne pouvant les citer tous, je rends particulièrement hommage à Laurence Rozé-Marchand avec qui j'ai partagé pendant des années le bureau A116 orange, pour l'ambiance amicale qu'elle a y su faire régner.

Table des matières

1 Introduction **5**

2 Théories linguistiques pour besoins applicatifs **11**
2.1 Quels besoins pour une application? 12
2.1.1 Ambiguïté lexicale 13
2.1.2 Formulations différentes d'un même concept 14
2.1.3 Phénomènes pris en compte 19
2.2 La sémantique différentielle 21
2.2.1 Description 22
2.2.2 Motivations 24
2.3 Le lexique génératif 26
2.3.1 Description 26
2.3.2 Motivations 30
2.4 Conclusion 31

3 Apprentissage de relations intracatégorielles basées sur la sémantique différentielle **33**
3.1 Méthodologie d'acquisition 35
3.1.1 Caractérisation des thèmes 36
3.1.2 Constitution et structuration de taxèmes 38
3.1.3 Bilan 41
3.2 Perfectionnement des étapes 42
3.2.1 Caractérisation des thèmes 42
3.2.2 Constitution et structuration de taxèmes 50
3.3 Conclusions 54

4 Apprentissage de relations nomino-verbales basées sur le Lexique génératif **57**
4.1 Le corpus et ses étiquetages 61
4.1.1 Étiquetage catégoriel 62
4.1.2 Étiquetage sémantique 62
4.2 Méthode d'apprentissage 64
4.2.1 Construction des exemples et des connaissances préalables . . 65
4.2.2 Production efficace d'hypothèses bien formées 66

4.3 Apprentissage basé sur les informations catégorielles et sémantiques 72
4.3.1 Validation théorique de l'apprentissage 72
4.3.2 Résultats et validation empirique 73
4.3.3 Comparaison avec des approches statistiques et syntaxiques . 75
4.3.4 Évaluation linguistique 78
4.4 Apprentissage sans prise en compte de l'étiquetage sémantique des noms 81
4.5 Bilan et discussions 84

5 Conclusions et perspectives **91**
5.1 Bilan 91
5.2 Perspectives 92

Chapitre 1

Introduction

Ce document est une synthèse d'une dizaine d'années de recherche dans le domaine du traitement automatique des langues (TAL), période débutant à la fin de ma thèse de doctorat au cours de laquelle je m'étais plus particulièrement focalisée sur des aspects syntaxiques de la langue, et pendant laquelle mes centres d'intérêt se sont orientés vers la sémantique.

Pendant ces dix ans, je me suis, dans un premier temps, intéressée à la modélisation du sens, en particulier des séquences complexes, cherchant à accroître les capacités de compréhension de ces composés en mettant au point un système de règles d'interprétation basées sur la sémantique des constituants simples et sur la forme de la structure complexe, et reposant sur des théories linguistiques. Mon activité a ensuite évolué vers l'acquisition d'éléments de sémantique lexicale en corpus, suivant en cela d'une part les besoins apparus lors de la constitution du modèle d'interprétation des composés et la mouvance générale qui se faisait jour en même temps dans la communauté TAL, mouvance liée à la disponibilité de quantités énormes de textes. C'est sur cet aspect apprentissage que se focalise ce document. J'ai en effet choisi de ne pas faire une rédaction chronologique de mes travaux, mais plutôt une synthèse de ceux-ci, mettant plus clairement en évidence les points fondamentaux de ma façon d'appréhender le TAL et mes contributions à ce domaine.

Mon travail de recherche se situe donc dans le cadre du TAL au sein duquel il porte plus précisément sur la sémantique lexicale – je résumerai par la suite sous la dénomination *TAL sémantique* ce domaine. Je m'intéresse au développement de méthodes d'apprentissage automatique de ressources lexicales sur corpus textuels et, plus précisément, de relations lexicales sémantiques permettant d'enrichir la description de mots, essentiellement de noms. Cet enrichissement n'est pas lié à un seul but lexicologique mais a pour objectif de répondre à un besoin applicatif mis au jour. Des théories linguistiques me servent alors de cadres pour déterminer les relations lexicales pertinentes afin de répondre à ce besoin, valider ce qui est acquis, voire mettre au point la méthode d'apprentissage nécessaire à cette acquisition.

Ces quelques lignes font ressortir les trois mots qui sont caractéristiques de mes

centres d'intérêt et de ma façon d'aborder le domaine du TAL sémantique : apprentissage, théories linguistiques et applications, termes qui reflètent les trois aspects auxquels je souhaite contribuer. Mon but n'est pas en tant que tel de développer des applications ; je considère une application TAL comme un champ d'étude dans lequel je cherche à déterminer où la prise en compte d'éléments de sémantique lexicale peut apporter un plus, c'est-à-dire résoudre un problème particulier. Ma démarche consiste alors, à partir du problème pointé, à trouver ce qui est nécessaire dans la description des mots d'un lexique pour répondre à ce type de besoin, à déterminer les cadres linguistiques pertinents pour « contrôler » cet enrichissement descriptif, et à mettre au point une méthode d'apprentissage sur corpus des éléments nécessaires, afin que, quel que soit le domaine de l'application, il soit possible de mettre en place les ressources lexicales utiles. Notons d'ailleurs que cette dernière phrase porte en elle un autre point fondamental pour moi : les ressources lexicales sémantiques peuvent certes aider à répondre à de nombreux besoins applicatifs mais à condition d'être adaptées au domaine de l'application. Si le thésaurus WordNet[1] a, par exemple, servi de base à plusieurs travaux applicatifs [Voo94, Fel98, Sme99], certains auteurs dont je partage l'avis réfutent la thèse de la pertinence de l'utilisation de telles ressources mutualisables, l'objection principale opposée à cette démarche étant qu'elle fait l'hypothèse qu'une ressource lexicale générale est valable hors contexte. De nombreuses études (*cf.* par exemple [BHNZ97]) ont montré que la définition des relations de proximité sémantique ne peut pas être menée hors domaine mais doit au contraire s'appuyer sur les caractéristiques du corpus de travail, et que ces relations doivent donc y être acquises.

Pour illustrer ma démarche et mon triple focus apprentissage - théories linguistiques - applications, prenons l'exemple d'une des applications d'accès au contenu de documents textuels que l'on retrouvera en fil rouge dans les chapitres suivants : la recherche d'information. Si l'on cherche à augmenter les taux de rappel et précision de systèmes de recherche d'information, on peut par exemple vouloir accéder à des variantes des éléments représentant les contenus des textes ou des requêtes[2]. Il faut donc disposer, au sein de la description des termes d'indexation du système étudié, de liens permettant de prendre en compte la diversité des formes sous lesquelles ces mots peuvent apparaître. À partir des termes d'indexation, on va donc définir un certain type de relations lexicales sémantiques pertinentes pour ce faire, par exemple établir des liens de synonymie afin d'être capable d'apparier une requête portant le mot *automobile* et un texte contenant le mot *voiture*. Des cadres théoriques peuvent alors donner les guides nécessaires pour établir les liens effectivement judicieux d'une part et une méthodologie d'apprentissage sur corpus de ces relations d'autre part.

Dans mes travaux, je m'intéresse ainsi à la mise au point de méthodes d'apprentissage sur corpus de deux familles de liens permettant d'enrichir la description lexicale de noms. D'une part, en me positionnant dans le cadre de la sémantique différentielle de Rastier [RCA94, Ras96], je cherche à apprendre, par des méthodes statistiques (en particulier de classification ascendante hiérarchique), des liens intra-

1. http://www.cogsci.princeton.edu/∽wn/
2. Cette discussion sera reprise et beaucoup plus développée au chapitre 2.

catégoriels (synonymie..., mais aussi d'autres liens plus « fins ») ; dans le cadre d'applications de type recherche d'information, l'acquisition de ces liens vise à répondre au problème évoqué au paragraphe précédent. D'autre part, en contrôlant leur pertinence grâce au formalisme du Lexique génératif de Pustejovsky [Pus95, BB01], j'acquiers par de l'apprentissage symbolique de type programmation logique inductive [MDR94] des liens transcatégoriels nomino-verbaux ; en recherche d'information, ces liens conduisent également à des reformulations intéressantes de termes d'indexation nominaux – ce qui constitue d'ailleurs une hypothèse très peu exploitée jusqu'ici – et permettent également de les désambiguïser. Ces deux théories de sémantique lexicale[3], l'une différentielle, l'autre descriptive, ont pour point commun de fonder très fortement les descriptions lexicales dans l'utilisation des mots en contexte. L'exposé de ces travaux d'apprentissage de relations lexicales sémantiques basées sur la sémantique différentielle et le Lexique génératif constitue le cœur de ce document. Tant l'utilisation peu habituelle d'une méthode d'apprentissage symbolique (*versus* une méthode statistique) que l'acquisition de liens intercatégoriels peu usités en reformulation (*versus* des liens nominaux intracatégoriels) me font considérer la seconde de ces études comme la plus originale – et également la plus aboutie – parmi mes contributions.

Mes travaux de recherche se situent donc à la croisée de diverses disciplines telles que le TAL, la lexicographie, la linguistique, l'apprentissage automatique et l'intelligence artificielle entre lesquelles ils cherchent à faire le lien. Si je tente de les caractériser par rapport aux nombreux travaux sur la sémantique lexicale, je dirais que contrairement aux recherches en apprentissage qui ne prennent pas en compte des questions de représentation linguistique d'une part, et à ceux qui portent sur des aspects linguistiques mais sans souci particulier d'acquisition d'autre part, je cherche à rendre des théories linguistiques et l'apprentissage compatibles, en ce sens que je suis guidée par des théories linguistiques et que je vise à acquérir des éléments dont la validité linguistique est attestée. Mes recherches sont également un lieu où le rapport entre théories linguistiques et applications est mis à l'épreuve. Ces deux points y cohabitent : une théorie linguistique a pour rôle de prédire ce qu'il faut acquérir, c'est-à-dire de déterminer, parmi l'ensemble des éléments sémantiques possibles, ceux qui peuvent influer sur l'application visée. Ainsi, dans le cadre de l'apprentissage de liens intercatégoriels permettant d'accéder à des reformulations de termes nominaux, le Lexique génératif indique certains liens nomino-verbaux sur lesquels il convient de se focaliser. De plus, les théories linguistiques décrivent des moyens pour accéder à ces éléments. Mes travaux posent donc la question de la façon dont des données pertinentes dans le cadre d'une théorie peuvent influencer les performances d'une application. Ce document n'apporte pas de réponse complète à ce problème – je débute seulement, par exemple, l'insertion effective de liens sémantiques appris au sein d'applications –, mais il est important de noter cette caractéristique de mes travaux. Enfin, si, comme je l'ai dit précédemment, mes recherches se caractérisent par les trois mots apprentissage, théories linguistiques et applications, je cherche non seulement à acquérir des éléments de sémantique

3. La sémantique différentielle ne se limite d'ailleurs pas au niveau lexical.

lexicale pertinents pour un besoin applicatif mais à contribuer effectivement à ces trois domaines très différents. Au fil de ce document, je montrerai en effet que pour répondre à des problèmes de recherche documentaire – en proposant le développement de méthodes d'acquisition de relations pour la désambiguïsation et l'expansion contrôlée des termes d'indexation –, les travaux menés portent sur les méthodes et algorithmes d'apprentissage (définition d'opérateurs de raffinement en programmation logique inductive ou traitement de problèmes liés à des matrices creuses en classification hiérarchique par exemple) et contribuent à la réflexion linguistique. Les travaux sur l'acquisition de liens nomino-verbaux basés sur le Lexique génératif contribuent, par exemple, à définir plus précisément le lien télique (but, fonction) dans ce formalisme ; de même, mes recherches visent à tester l'implémentabilité de théories linguistiques et donc la possibilité de développement « à grande échelle » de lexiques basés sur leurs principes, utilisables dans des cadres applicatifs.

Le plan de ce document reprend précisément les idées énoncées dans cette introduction. Le chapitre 2, intitulé *Théories linguistiques pour besoins applicatifs*, donne en effet les bases des théories linguistiques qui me servent de cadre pour définir les éléments de sémantique lexicale permettant d'enrichir la description de noms dans une double optique de désambiguïsation et de traitement de variantes sémantiques intra- et intercatégorielles, et susceptibles d'être utilisés au sein de systèmes de recherche d'information pour améliorer les possibilités d'appariement de requêtes et de textes et accroître leur précision. Les chapitres 3 et 4 décrivent quant à eux les méthodes d'apprentissage en corpus de ces éléments et les résultats concernant l'acquisition de ces liens intra- et intercatégoriels définis dans les cadres linguistiques précédents. Ils font également le bilan de mes contributions dans les trois domaines fondamentaux que sont pour moi l'apprentissage, la linguistique et les applications (ici la recherche d'information essentiellement). Enfin, le chapitre 5 présente des réflexions et perspectives de travail.

Avant de passer à l'exposé, je voudrais signaler que le travail décrit dans ce document a, bien évidemment, bénéficié de collaborations multiples, en particulier localement de celles de membres des ex-équipes Repco et Aïda et de l'équipe TexMex de l'Irisa, permanents, thésards ou stagiaires. Le « je » que *je* me suis autorisée à employer au cours de cette introduction pour décrire ma façon d'appréhender le TAL sémantique sera donc remplacé par le « nous » plus traditionnel de la rédaction scientifique qui reflète réellement le travail commun réalisé. Les personnes ayant effectivement participé aux recherches décrites seront associées au fil du texte aux chapitres ou sections adéquates. Cependant, ce texte fait l'impasse sur une partie des travaux que j'ai effectués pendant la même période sur l'intégration du modèle d'interprétation de composés au sein d'un tuteur destiné aux étudiants francophones de l'anglais ; ceux-ci ont donné lieu à la thèse de doctorat de Frédéric Danna que j'ai encadrée[4] et qui a été soutenue en janvier 1997, ainsi qu'aux stages de DEA (outre celui de Frédéric) de Christine Largouët, David Galic et Guillaume Maingourd. Ils ont également conduit à une collaboration locale avec Dominique Py et extérieure

4. Thèse dirigée par Marie-Odile Cordier.

avec Paul Boucher (Université de Nantes). Ne pouvant les mentionner au fil du document, je tiens à le faire ici tout en les remerciant.

Chapitre 2

Théories linguistiques pour besoins applicatifs

Nous avons, en introduction, esquissé les grandes lignes de nos travaux qui consistent, en se situant « sous le contrôle » d'un cadre formel linguistique, à mettre au point des méthodes d'apprentissage automatique en corpus de relations lexicales sémantiques, susceptibles d'être utilisées dans le cadre d'applications TAL pour répondre à un besoin mis au jour. Ce second chapitre, que nous aurions également pu intituler « Sémantique et applications TAL dans un cadre formel linguistique », ne porte pas sur l'aspect apprentissage qui fera l'objet des chapitres suivants, mais concerne les autres points cités dans cette première phrase. Son objectif est double. D'une part, en prenant pour support illustratif les applications TAL qui nécessitent un accès au contenu de documents textuels, nous montrons des besoins auxquels un apport de connaissances sémantiques peut répondre, et définissons par là même le type d'éléments sémantiques lexicaux qu'il est donc nécessaire d'acquérir. Ce premier point, qui, outre un éclairage général des problèmes, met en lumière les connaissances lexicales que *nous* voulons apprendre sur corpus, fait l'objet de la section 2.1.

Le second objectif de ce chapitre est de présenter les principes des théories linguistiques dans le contexte desquelles nous nous situons, afin d'avoir un cadre formel d'acquisition des relations lexicales sémantiques[1] pouvant répondre aux besoins applicatifs pointés précédemment. Le rôle de ces théories est de définir une méthodologie d'apprentissage de l'information pertinente et d'offrir une validation des éléments acquis. Nous présentons successivement les deux cadres théoriques qui concernent plus particulièrement nos travaux. La section 2.2 décrit ainsi succinctement les points pertinents pour nous de la *sémantique différentielle* de Rastier [RCA94, Ras96], qui définit des relations lexicales sémantiques de similarité et différenciation entre mots au sein d'une même catégorie grammaticale. La section 2.3 est, quant à elle, dédiée à l'explicitation du *Lexique génératif* de Pustejovsky

1. Nous parlerons par la suite indifféremment d'acquisition d'éléments lexicaux sémantiques, d'acquisition de lexiques sémantiques, ou d'acquisition de relations lexicales sémantiques.

[Pus95, BB01], qui associe à la définition lexicale des mots – et en particulier des noms sur lesquels nous nous focalisons – divers prédicats verbaux. Dans ces deux sections, nous nous en tenons à une présentation non exhaustive des théories, plutôt axée sur les points d'intérêt qui permettront de comprendre en particulier *et* la terminologie, *et* la méthodologie d'acquisition des éléments lexicaux sémantiques dans les deux chapitres suivants.

2.1 Quels besoins pour une application?

Avec le développement des réseaux internes et mondiaux (Intranet, Internet...), la quantité de textes électroniques disponibles s'est considérablement accrue. Des applications ont alors vu le jour ou ont pris une place beaucoup plus importante. Si en 1993, parmi les trois tâches qui pour eux prévalaient ou allaient prévaloir dans les travaux liés à l'interprétation des textes, Jacobs et Rau [JR93] distinguaient clairement la recherche d'information et la catégorisation de textes, on regroupe actuellement volontiers sous des vocables plus généralistes (*gestion d'information basée sur le contenu* pour [Sme99], *recherche d'information* pour [Str99] par exemple) l'ensemble des recherches portant sur la classification, la catégorisation, le filtrage, l'indexation et la recherche... dans de grands volumes de données essentiellement textuelles.

Pour illustrer notre propos, nous prenons ici pour représentant de ces travaux la recherche d'information (RI) dans son acception recherche documentaire, qui consiste à offrir à un utilisateur interrogeant une base de textes ceux qui satisfont au mieux sa requête. Dans les systèmes de recherche d'information (SRI), les documents de la base textuelle et les requêtes des usagers sont, en règle générale, représentés par des mots qui, dans le cas des systèmes automatiques, sont très fréquemment des noms[2] (N), essentiellement simples, extraits automatiquement de ces documents et requêtes. Un appariement entre mots est donc utilisé pour déterminer les textes à proposer en réponse à une interrogation. Ceci pose deux problèmes d'ordre sémantique :

- l'ambiguïté des mots : cette ambiguïté, qui existe même si l'on se limite à une collection de documents très ciblée comme le signale Krovetz [Kro97], peut conduire à proposer des textes non pertinents pour une requête ;
- les formulations différentes d'un même concept : un document intéressant peut contenir des mots autres que la requête : synonymes, formes morphologiquement différentes...

Il convient donc de trouver des moyens d'enrichir la description des termes d'indexation pour être capable de sélectionner automatiquement le sens pertinent d'un mot parmi ses diverses significations possibles et d'en effectuer une expansion contrôlée afin d'accéder à des formulations différentes du concept exprimé dans une requête (voire un document). De manière plus générale, en sortant du cadre de la RI, ces

2. Éventuellement des termes, ceux-ci étant de bons descripteurs de l'information contenue dans un document [JKT97].

deux phénomènes constituant des problèmes linguistiques en soi et nous intéressant en tant que tels, il nous faut réfléchir à l'enrichissement de la description lexicale des noms dans la double optique de désambiguïsation et de traitement des différentes variantes sous lesquelles ils peuvent apparaître.

Nos recherches portent donc sur la mise au point d'une méthodologie d'apprentissage sur corpus d'éléments lexicaux susceptibles de répondre à ce double besoin et, pour ce faire, nous choisissons des cadres linguistiques théoriques pour nous guider vers cet objectif d'acquisition. Mais avant de présenter les cadres que nous avons retenus et les raisons de ces choix, nous allons successivement faire une rapide présentation non exhaustive des travaux qui, tant dans le cadre de la RI que dans celui du TAL ou autres domaines « proches », se sont intéressés à traiter ces deux problèmes d'ambiguïté lexicale et de variations. Nous mentionnons ensuite, parmi l'ensemble de tous les points évoqués dans ce bref tour d'horizon, ceux auxquels nous nous intéressons plus particulièrement.

2.1.1 Ambiguïté lexicale

Le traitement de l'ambiguïté et de la sélection du sens d'une occurrence d'un mot donnée est, comme nous venons de le dire, un problème linguistique général, même si nous étudions ici essentiellement les difficultés qu'il implique en RI. De nombreux travaux sur la désambiguïsation du sens des mots (*word sense disambiguation*, WSD) [IV98, Yar95] ont d'ailleurs été dédiés à ce point en dehors de tout cadre applicatif. C'est par exemple le cas de ceux rassemblés au cours des campagnes Senseval pour l'anglais ou Romanseval pour les langues romanes, compétition à laquelle nous avons d'ailleurs participé en 1999 (*cf.* [CHU00] pour une présentation des travaux de cette année-là). Dans ces campagnes, l'ensemble des sens possibles des mots cibles de la désambiguïsation sont connus et il convient, pour chaque occurrence en contexte, de déterminer celui qui est concerné.

Dans le cadre de la RI, plusieurs auteurs ont pointé ce problème de l'homonymie et de la polysémie et des chutes de performances des systèmes que cela peut entraîner (par exemple [Kro97, Voo98] et [Sme99]). Le recours à des termes d'indexation complexes (indexation syntagmatique) est une des solutions proposées [Fag87, Sal89, GGHR00] pour réduire l'ambiguïté, un mot ayant en règle générale un sens plus précis au sein d'une structure composée, comme par exemple *cours* dans une structure *N de N* telle que *cours de l'Euro* (versus *cours de mathématiques*). C'est aussi celle que nous avons retenue dans [FS99]. De même, un terme composé de la forme A [3] N peut être analysé comme une dépendance tête-modifieur, et on peut alors vouloir reconnaître la présence de cette dépendance entre deux mêmes éléments simples dans un document et une requête, quelle que soit la forme de surface dans laquelle elle apparaît, et éviter ainsi certaines ambiguïtés.

Plus généralement, l'idée que des mots cooccurrant dans un contexte codéterminent leurs sens appropriés même si chacun est ambigu, proche d'ailleurs de la notion d'isotopie de Rastier [RCA94, Ras96] dont nous parlerons brièvement en section 2.2, est utilisée, avec des fortunes diverses, dans [Voo98] et [RBC00]. Dans

3. Adjectif.

le premier travail, les mots environnant une occurrence d'un mot m dans un texte ou une requête servent à favoriser le choix d'un sens de m dans la base lexicale WordNet. Cependant, l'introduction dans un SRI des mécanismes de désambiguïsation proposés ne conduit pas à des résultats extrêmement probants, en particulier pour traiter les questions courtes. Le second travail, dans lequel la représentation d'un document prend en compte les fréquences de cooccurrence de ses unités linguistiques avec les termes d'indexation choisis – C_{ij}, nombre de cooccurrences du mot i avec le terme d'indexation j, étant alors vu comme un estimateur de la probabilité que l'unité linguistique i exprime le sens j – conduit, quant à lui, à des améliorations des performances du SRI testé.

Pour terminer ce rapide tour d'horizon du traitement de l'ambiguïté, en particulier en RI, nous mentionnons le travail de Grefenstette [Gre97] qui prône un enrichissement de la représentation des N présents dans des requêtes courtes par les éléments des structures linguistiques dans lesquelles ces N apparaissent dans la collection interrogée. Il ne s'agit pas d'expansion de requêtes mais d'une forme d'ergonomie linguistique de la RI, où l'utilisateur se voit proposer des liens syntagmatiques, souvent nomino-verbaux (sujet-verbe, verbe-objet...), qui aident à préciser et à désambiguïser les N, et parmi lesquels il effectue un choix pour permettre au SRI de lui répondre de manière plus satisfaisante. Par exemple, Grefenstette montre qu'un moyen de caractériser sémantiquement un nom comme *research* est d'extraire l'ensemble des verbes utilisés avec lui, de manière à recenser ce que la recherche permet de faire (*research show*, *research reveal*...) et ce qui est fait en direction de la recherche (*do research*, *support research*...).

2.1.2 Formulations différentes d'un même concept

Le second problème récurrent des SRI que nous avons noté concerne les formulations différentes d'un même concept et l'influence de ce phénomène sur leurs performances (le rappel en particulier). Il est nécessaire de trouver un moyen d'introduire de la flexibilité dans la procédure d'appariement des représentations des documents et des requêtes, afin que des réalisations linguistiques différentes mais portant le même contenu informationnel puissent être considérées comme équivalentes. Là encore, le phénomène des variantes sous lesquelles une idée peut apparaître n'est pas un problème limité à la seule RI, mais est un sujet d'étude linguistique beaucoup plus général qui a donné lieu à de très nombreux travaux en linguistique, TAL, terminologie, linguistique de corpus... (que nous regroupons sous l'appellation TAL par la suite) ; seuls les résultats de certains d'entre eux ont été intégrés à des SRI, parfois uniquement partiellement, et avec des bénéfices plus ou moins probants qui soulignent la précaution qu'il convient de prendre dans l'application du TAL à la RI[4]. D'ailleurs, dans leur grande majorité, les systèmes actuels de RI n'exploitent encore pas ou peu d'informations de nature linguistique [Sme99]. Ceux qui y ont

4. Nous sommes bien consciente de la distance actuelle entre les « éléments » que l'on peut acquérir sur corpus par des méthodes de TAL et d'apprentissage automatique et leur application directe en RI, comme le signalent par exemple [Str99, SJ99] et [Jac00], et nous aurons l'occasion de discuter de ce point par la suite.

néanmoins recours se donnent, comme nous l'avons déjà abordé en sous-section 2.1.1 et allons le voir ici, deux objectifs :

- Le premier concerne la définition de descripteurs de contenu correctement discriminants et non ambigus (index complexes, index structurés syntaxiquement [SJ99]).
- Le second vise, quant à lui, essentiellement l'augmentation des possibilités d'appariement requêtes/documents, surtout en exploitant des relations synonymiques ou hyperonymiques répertoriées dans des bases lexicales [DRF89, Sme99].

Nous étudions brièvement certains des travaux de TAL[5] liés aux phénomènes de variations morphologiques, (morpho-)syntaxiques et sémantiques des mots (voir par exemple [Dai02] pour une typologie plus complète des variations), et présentons les aspects de ces variations pris en compte en RI.

Variations morphologiques De nombreux travaux ont été menés en TAL pour permettre de reconnaître ou générer les diverses formes des mots d'une même famille morphologique ; le chapitre de livre [DFS02] co-écrit avec Béatrice Daille et Cécile Fabre en montre d'ailleurs plusieurs et décrit plus en détail les recherches citées ici. Une analyse morphologique, fondée sur l'utilisation d'une base lexicale[6] ou d'un analyseur[7], permet d'obtenir le lemme ou la racine d'un mot étudié, ainsi que les divers traits morphologiques (genre, personne...) ou les affixes qui le caractérisent. Outre celles portant sur le développement et l'utilisation de ces outils d'analyse, des recherches ont également été menées dans le but de constituer automatiquement des familles morphologiques avec des connaissances linguistiques initiales nulles ou minimales, telles que, par exemple, [GZ99] et [Jac97].

En RI, l'analyse morphologique (et en particulier la racinisation intégrée massivement dans les SRI portant sur l'anglais) est utilisée pour regrouper des termes appartenant à une même famille morphologique. Ceci permet soit de représenter les mots de la même famille par un même terme d'indexation ou bien d'étendre un mot d'une requête par sa famille morphologique. L'influence de l'analyse morphologique en RI a été évaluée dans divers systèmes pour l'anglais, ce qui a conduit à quelques débats quant à son intérêt pour la tâche de recherche : [LPTW81] et [Har91] concluent que la racinisation n'améliore pas les résultats – et en particulier que des procédures sophistiquées d'analyse donnent des résultats similaires à des solutions très basiques –, alors que, par exemple, [XC98] et [SLWPC99] parviennent à la conclusion inverse. D'autres expériences ont montré que la racinisation ou la lemmatisation amélioraient significativement les performances pour les langues à caractère morphologique plus fort comme l'italien ou le français [GGS97, JT99, GGHR00], ce dernier article, par exemple, mettant en particulier l'accent sur le fait que l'analyse morphologique favorise le rappel mais également la précision.

5. Nous n'avons pas ici la volonté de faire une présentation exhaustive des travaux portant sur les divers points mentionnés.

6. Par exemple, Delas [CS89], Celex [Bur90]...

7. Des racineurs, tels que ceux développés par Lovins [Lov68] et Porter [Por80], ou des outils basés sur des techniques d'analyse morphologique plus sophistiquées [Ant90, Nam00].

Variations syntaxiques et morpho-syntaxiques Les variations syntaxiques des termes complexes (*gene expression / gene amplification and expression*), de même que les variations combinant transformations morphologiques et syntaxiques, ont donné lieu à de multiples études, en particulier dans le cadre de travaux d'acquisition de terminologie [Jac01, BJL01].

Jacquemin [Jac96] définit ainsi une variation morpho-syntaxique comme une variante vérifiant quatre conditions : 1- les mots pleins (N, V, A) du terme initial sont présents ; 2- ces mots pleins peuvent subir des modifications morphologiques flexionnelles ou dérivationnelles ; 3- l'ordre des mots peut être altéré et des mots peuvent être insérés, mais la relation de dépendance entre les mots pleins du terme initial doit être maintenue dans la variante ; 4- la variante ne doit pas contenir la chaîne du terme initial aux flexions près.

Le logiciel `Faster` développé pour le repérage des variantes de termes propose ainsi des métarègles de réécriture d'un terme pour obtenir ses variantes correctes. Par exemple, la métarègle :

N1 P2 N3 → V1 (Av? (P[8]? D[9] | P) A?) N3
<V1 der ref> = <N1 ref>

peut s'appliquer sur une structure $N\,P\,N$ et reconnaître la variante indiquée à droite dans laquelle le verbe *V1* doit avoir un lien morphologique avec le premier nom *N1*. Quelques catégories grammaticales peuvent s'insérer, *?* marquant l'optionnalité et | la disjonction. Par exemple, cette métarègle reconnaît *stabiliser leur prix* comme variante de *stabilisation de prix*.

`Faster` a été utilisé à partir d'une liste de termes initiaux pour repérer les variantes présentes dans des corpus textuels [JT99, JKT97]. L'utilisation de cet outil permet ainsi d'identifier 30% d'occurrences de termes supplémentaires, et donc de fournir une indexation à couverture beaucoup plus large.

De même, le système `Acabit` [Dai00] d'acquisition de termes est capable de prendre en compte des variantes morpho-syntaxiques des structures de base ($N\,A$, *N (P (D)) N*, *N à* V_{inf} [10]) qu'il reconnaît, comme par exemple le couple *conquête spatiale / conquête de l'espace*.

Dans le cadre de la RI, plus que l'intégration effective des travaux de terminologie présentés ci-dessus, la prise en compte des variantes syntaxiques se fait par la production d'index structurés via une analyse syntaxique, ou par expansion d'une dépendance syntaxique d'une requête par sa classe de formes possibles. Plusieurs études ([SJ99], [SLWPC99]...) utilisent ainsi des index structurés, obtenus par une analyse soit des seules questions, soit des questions et des documents, et une normalisation de ces descriptions afin de permettre un appariement entre index superposables au niveau de la structure et proches au niveau du concept abordé, en passant outre les variations syntaxiques sous lesquelles les mêmes termes composés peuvent apparaître (*cf.* l'exemple *information retrieval, retrieval of information, information that is retrieved...* dans [SLWPC99]). [SLWPC99] montre le gain apporté par des index complexes structurés de type tête-modifieur, surtout dans le cas de

8. Préposition.
9. Déterminant.
10. Verbe à l'infinitif.

requêtes longues. Spärck Jones et Tait [SJT84] déterminent la structure propositionnelle des questions dont on peut extraire les termes complexes intéressants et utilisent des classes d'équivalence permettant de générer les diverses alternatives qui peuvent correspondre à la même relation syntaxique dans un texte. [Sme99] utilise, quant à lui, des représentations arborées des requêtes et des documents qui, si elles fonctionnent bien pour des appariements syntagmes à syntagmes, sont d'un intérêt beaucoup moins convaincant lorsqu'elles sont insérées dans un SRI.

Variations sémantiques Comme nous l'avons vu, un même contenu peut être exprimé de manières différentes, dans différentes configurations syntaxiques, avec différents mots. Cependant, le diagnostic de paraphrase, dès lors qu'il dépasse le cadre strict de la transformation (morpho-)syntaxique, est extrêmement difficile à contrôler et requiert des informations linguistiques riches. Ainsi, si l'on veut pouvoir apparier *vélo* et *bicyclette*, ou encore *professeur de mathématiques* et *enseigner les mathématiques*, il est nécessaire de posséder une base de connaissances lexicales contenant ces relations sémantiques entre mots, que celles-ci soit intracatégorielles (entre deux mots de la même catégorie grammaticale comme le premier exemple) ou intercatégorielles (entre mots de catégories différentes comme le lien sémantique entre le N *professeur* et le V *enseigner*). Les relations sémantiques sont généralement classées en deux familles : les relations syntagmatiques et les relations paradigmatiques. Les premières dénotent les capacités d'association d'un mot, le contexte qu'il sélectionne ; c'est par exemple, pour un verbe, sa structure argumentale (le nombre et le type de ses arguments), ou, pour un nom, les collocations auxquelles il participe ou les verbes qui, comme dans l'exemple ci-dessus, révèlent des aspects de sa signification. Les secondes regroupent des mots d'un même paradigme à fonctionnements quasi similaires, que ce soit en constituant des classes sémantiques ou en définissant des liens de synonymie, d'hyperonymie...

Dans le cadre de la RI, la démarche la plus souvent adoptée pour disposer de ces liens sémantiques entre mots consiste à recourir à une base de connaissances linguistiques regroupant les mots sémantiquement proches, et structurée, en règle générale, selon des relations hyperonymiques ou synonymiques. Les termes d'indexation d'une requête peuvent alors être automatiquement propagés en suivant les liens exprimés dans la base lexicale, de manière à disposer d'une description étendue de la requête. C'est par exemple l'option choisie dans [GLO93] pour la consultation du Minitel en français. Dans ce type d'approche, le poids d'un document est affaibli en fonction du lien suivi dans la base lexicale et de la distance entre le mot dans la requête et dans le document.

On connaît le coût de construction de telles ressources, qui amène généralement ceux qui adoptent cette approche à plaider pour l'utilisation de ressources générales, mutualisables, dont WordNet constitue le modèle [Sme99, Fel98, Voo94]. Le gain apporté par le recours à de telles ressources n'a jusqu'à présent pas réellement été démontré. Certaines expériences tendent même à invalider cette approche. Voorhees [Voo94] souligne, par exemple, que, si la sélection manuelle des *synsets*[11] pour étendre les requêtes, en particulier courtes, peut apporter un plus, l'automa-

11. Ensembles de mots exprimant un concept.

tisation problématique de ce choix rend les résultats très peu probants. Smeaton [Sme99], qui utilise WordNet pour calculer une distance sémantique entre mots en se basant sur une méthode proposée dans [Res95a, Res95b], souligne lui aussi ce problème, qui montre que les deux difficultés d'ordre sémantique (ambiguïté et formulations diverses de concepts) que nous avons pointées en les dissociant pour des besoins rédactionnels sont bien évidemment beaucoup plus imbriquées dans la réalité. Comme nous l'avons déjà signalé en introduction, certains auteurs, dont nous partageons les idées, réfutent d'ailleurs la thèse de la pertinence de l'utilisation de telles ressources, l'objection principale étant que cette démarche fait l'hypothèse qu'une ressource lexicale générale est valable hors contexte. Or, de nombreux travaux (par exemple, [BHNZ97, PW95]) ont montré que la définition des relations de proximité sémantique ne peut pas être menée hors domaine mais doit au contraire s'appuyer sur les caractéristiques du corpus de travail. De façon plus générale, l'utilisation systématique de WordNet dans le domaine du traitement automatique des langues est sujette à caution : dans quelle mesure un modèle sémantique conçu *a priori* s'avère-t-il adéquat pour représenter le fonctionnement de domaines particuliers ? Cette question de fond n'est pas toujours soulevée, et n'est en tous cas pas encore résolue, par ceux qui l'utilisent. En outre, étendre une requête consiste précisément à tenter de la rapprocher des documents qu'elle cherche à explorer, en d'autres termes, à l'ancrer dans les mots réellement utilisés dans le corpus des textes de l'application traitée.

Pour pallier ce problème, outre l'option suivie par certains de spécialiser grâce à des textes du domaine une base lexicale générale [Bui97, VFP91], la solution consiste à acquérir, à partir de corpus, l'intégralité des connaissances lexicales sémantiques requises. De très nombreux travaux (*cf.* [Gre94b], par exemple, ou [HNS97] et [PS97] pour des états de l'art du domaine) ont déjà été réalisés sur le sujet, essentiellement dans le cadre de l'apprentissage statistique, même si depuis quelques années, des recherches sur l'acquisition en corpus de lexiques ou relations sémantiques ont également vu le jour dans le cadre de l'apprentissage symbolique [WRS96]. Ces recherches visent à extraire des informations syntagmatiques et paradigmatiques sur les unités lexicales, étudiant respectivement les mots qui apparaissent dans les mêmes fenêtres ou les mêmes contextes syntaxiques que l'unité considérée (affinités du premier ordre pour reprendre les termes de Grefenstette [Gre94a]), ou les mots qui génèrent les mêmes contextes que le mot cible (affinités du second ordre). Par exemple, [BC97] et [FN99] tentent d'apprendre automatiquement des structures argumentales et des restrictions sélectionnelles ; [GT95] acquièrent des verbes supports de nominalisations ; [Aga95] et [BHNZ97] construisent des classes sémantiques ; [Hea92, Hea98] et [Mor97] se focalisent sur un type particulier de relation lexicale, telle que l'hyperonymie ; [Gre94b] vise de son côté l'obtention de représentations lexicales sémantiques plus complètes.

Ces recherches se situent généralement dans le cadre de la linguistique harrissienne [HGR+89] qui pose l'hypothèse que des regroupements de mots opérables sur la base de fonctionnements linguistiques communs en corpus permettent l'identification et la structuration de catégories conceptuelles, soit, en d'autres termes, qu'il est possible de mettre en évidence, à partir d'une analyse distributionnelle de contextes

rendus élémentaires, les classes de concepts et les relations d'un sous-langage lié à un domaine d'activité. Ainsi, parmi les nombreux travaux sur l'apprentissage de relations paradigmatiques par des méthodes statistiques, dont le but est de faire émerger des mots qui ont des comportements similaires, ceux qui portent sur l'acquisition automatique de classes sémantiques suivent, en général, une méthodologie proche, décomposable en trois phases [Gre94b]. La première phase concerne l'extraction des cooccurrents d'un mot, c'est-à-dire des mots qui apparaissent par exemple avec lui soit dans un contexte syntaxique donné, soit dans une fenêtre de m mots (n mots avant le mot étudié, m-n mots après, par exemple). La seconde phase associe à chaque mot ses cooccurrents et met en évidence la proximité ou la distance des mots deux à deux en fonction des cooccurrents qu'ils partagent ou non ; là aussi, la mesure de proximité/distance entre les mots varie selon les travaux et est souvent basée sur un calcul statistique tel que la distance euclidienne, l'information mutuelle... Enfin, la dernière phase consiste à produire des classes en fonction des plus ou moins grandes proximités entre les mots. Cependant, la notion de similarité de comportement manipulée par ces méthodes est à comprendre au sens de Cruse [Cru86], c'est-à-dire que deux mots sont similaires s'ils sont substituables dans un même contexte. Les relations entre les mots ainsi regroupés peuvent alors être diverses et il convient souvent de les interpréter ou de les adapter manuellement.

2.1.3 Phénomènes pris en compte

Parmi les diverses formes de variations possibles des noms, nos travaux portent principalement sur la prise en compte des variations d'ordre sémantique. Nous nous intéressons d'une part à l'acquisition de relations lexicales paradigmatiques. Comme nous allons le voir, nous nous appuyons pour ce faire sur la théorie de la *sémantique différentielle* de Rastier qui a, entre autres, pour avantage de nous offrir des pistes pour développer une méthodologie d'apprentissage de divers éléments. Notre objectif n'est pas uniquement d'acquérir des liens intracatégoriels (et plus exactement entre noms) de type synonymique ou hyperonymique exploitables, par exemple, en RI, mais également de tenter de pointer automatiquement au sein de classes de quasi synonymes, acquises par des méthodes d'analyse des données, les relations fines de sens qui regroupent et différencient ces mots et qui sont obtenues par des traitements manuels dans [FHL97] par exemple.

D'autre part, nous nous focalisons également sur l'apprentissage de liens nomino-verbaux. Nous considérons en effet que la prise en compte du phénomène de variation sémantique des noms ne doit pas se limiter aux seules variantes exploitant des relations intracatégorielles, mais que la force du lien nomino-verbal doit elle aussi être mise à contribution. Ceci est particulièrement vrai dans le cadre de la RI dont les besoins nous ont servi de fil rouge tout au long de cette section. Nous pensons que pour répondre au besoin de reformulation des SRI, on ne peut se contenter d'acquérir et exploiter des relations paradigmatiques intracatégorielles dans lesquelles le seul N joue une place prépondérante, et estimons qu'il manque à l'approche par thésaurus une réflexion linguistique préalable concernant le fonctionnement sémantique des descripteurs. Elle mobilise en effet exclusivement les relations lexicales

traditionnelles (hyperonymie, synonymie), mais cette option témoigne d'une vision très cloisonnée du lexique. Ainsi Smeaton [Sme99] déclare n'exploiter de WordNet que les noms, ceux-ci étant les principaux détenteurs du contenu des textes. Or, plusieurs travaux, tant en terminologie qu'en analyse et typologie des textes [BC99, KK98], ont montré l'importance des verbes. S'il est prouvé que les groupes nominaux constituent le principal mode d'expression des descripteurs, l'apport sémantique du verbe ne doit donc pas être négligé pour réaliser l'enrichissement et la reformulation des termes d'indexation, et plus généralement, hors cadre applicatif, pour traiter de la variation sémantique des noms.

Nous avons déjà eu l'occasion de pointer l'importance du lien nomino-verbal dans le cadre de travaux sur la modélisation de la sémantique des groupes nominaux de forme *N N* en anglais et *N Prép (Dét) N* en français, qui ont donné lieu à la thèse de Cécile Fabre [Fab96] que j'ai encadrée[12]. Nous avons montré qu'une part essentielle du contenu de ces structures renvoie à des informations de nature prédicative, exprimables à l'aide d'un verbe. Cela est vrai bien sûr pour les séquences dont le nom tête est un nom morphologiquement dérivé d'un verbe (*interpréteur de commandes*), mais également pour celles où le lien avec un verbe n'est pas explicite (*parc à munitions* - *entreposer*, *stocker* ; *magasin de disques* - *vendre*). Ce phénomène prouve la force de l'association verbo-nominale, puisque l'information prédicative est associée au nom au point de pouvoir être non explicite dans ce type de structures. La mise en évidence d'un schéma événementiel attaché au groupe nominal offre des possibilités étendues de reformulation sur la base de liens intercatégoriels nom-verbe. Ainsi, en se basant sur le lien fonctionnel entre *magasin* et *vendre*, on peut accéder à une variante telle que *vendre des disques* à partir de *magasin de disques*. Nous voulons maintenant nous focaliser sur l'acquisition de ces liens nom-verbe.

Plusieurs autres études notent également l'intérêt de ces relations. Ainsi, contrairement à l'approche choisie dans WordNet, les concepteurs d'EuroWordNet[13] ont ajouté certains liens intercatégoriels dans leurs réseaux sémantiques [Vos98] (liens entre des concepts lexicalisés par différentes catégories). Par ailleurs, [FJ00] décrit une expérience qui vise à prendre en compte la variation nomino-verbale des termes afin d'exploiter le lien entre des termes nominaux (ex : *méthode d'obtention*) et des formulations verbales proches (ex : *obtenues par d'autres méthodes*). Ce travail constitue donc une première étape vers la prise en compte de critères de reformulation sémantique pour exploiter la relation nom-verbe. Le but de cette expérience est d'augmenter l'ensemble des catégories de variation terminologique traitées par `Faster` [JKT97] dont nous avons parlé plus haut (*cf.* page 16). Dans cette expérience, seule la relation nom-verbe validée par un lien morphologique est prise en compte. Elle est contrôlée à l'aide de quelques informations linguistiques notant, par exemple le caractère transitif ou non du verbe et la nature morphologique du nom (déverbal ou non), dont l'exploitation conduit à assurer que la relation argumentale entre les deux termes pleins dans la forme initiale est maintenue dans la variante verbale. Ce travail permet, par rapport à la première version de métarègles de trans-

12. Thèse effectuée sous la direction de Marie-Odile Cordier.
13. http://www.illc.uva.nl/EuroWordNet/

formation décrites dans [JT99], dont l'intérêt pour le repérage de reformulations de termes est attestée, d'accroître de 30% la précision des variantes détectées (faible baisse de 10% du rappel), et confirme l'intérêt des verbes et de leur sémantique pour l'analyse des textes. Nous proposons, quant à nous, d'étendre le traitement de cette relation intercatégorielle au cas d'associations nom-verbe sans lien morphologique.

Nous avons également déjà cité l'importance de ce lien nomino-verbal pour Grefenstette [Gre97] qui considère qu'il peut aider à préciser et à désambiguïser les noms contenus dans des requêtes courtes. Nous nous intéressons donc à la fois au lien nom-verbe dans une optique de traitement de la variation sémantique mais également dans une optique de désambiguïsation, et proposons des moyens de systématiser la proposition de Grefenstette.

Pour apprendre sur corpus ces liens sémantiques nomino-verbaux, nous devons donc choisir un moyen de déterminer les paires nom-verbe effectivement pertinentes du point de vue de l'enrichissement des noms. Par exemple, nous voulons trouver un cadre théorique définissant une relation (but ou fonction) entre le N *jaugeur* et le V *mesurer* qui nous permette d'accéder à l'extension intercatégorielle *jaugeur de carburant – mesurer du carburant*. Nous avons choisi le *Lexique génératif* de Pustejovsky [Pus95, BB01] pour définir ces paires (hypothèse justifiée en section 2.3) et nous exploitons la puissance de la programmation logique inductive [MDR94] pour acquérir ces liens nomino-verbaux sur corpus.

Nous venons de citer les deux théories linguistiques qui servent de cadre formel à nos travaux d'apprentissage en corpus de relations lexicales sémantiques permettant d'enrichir la description des noms dans une double optique de traitement de la variation sémantique et de désambiguïsation. Les deux sections suivantes en présentent succinctement les grandes lignes, en insistant sur les aspects adéquats pour notre propos, et elles justifient nos choix.

2.2 La sémantique différentielle

Dans cette section, nous exposons certains aspects clés de la sémantique différentielle (SD) de Rastier [14], théorie qui, comme nous le verrons, nous sert de cadre formel pour définir une méthodologie d'acquisition en corpus, sans pré-connaissances, de lexiques sémantiques basés sur des liens intracatégoriels de type synonymique, antonymique... mais également sur des relations sémantiques plus fines de distinction entre mots d'un même paradigme. Nous ne faisons ici qu'une présentation très partielle de cette théorie, en nous limitant le plus souvent aux points nécessaires à la compréhension du chapitre suivant. Il convient donc de se tourner vers les textes de Rastier [Ras96, Ras95, RCA94] pour en avoir une vue plus générale et plus précise, textes qui servent d'ailleurs de base à notre exposé. Des présentations intéressantes de cette théorie sont également disponibles dans [Beu98] ou [Tan97] par exemple.

14. Cette théorie est aussi nommée sémantique interprétative, mais nous privilégions ici l'appellation sémantique différentielle car c'est cet aspect différentiel que nous étudions plus particulièrement.

Nous terminons cette section en expliquant les raisons qui motivent la sélection de ce cadre linguistique pour nos travaux.

2.2.1 Description

La SD est issue d'une double influence : d'une part la linguistique structurale et ses idées reprises de Saussure, d'autre part l'herméneutique et ses théories de l'interprétation des textes (d'auteurs tels que Schleiermacher...). À la première elle emprunte les notions de valeur et de langue vue comme un système différentiel de signes, ou plus précisément de signifiés. Un signifié linguistique s'analyse donc en relations d'opposition avec les autres signifiés ; les traits relationnels qui le composent et qui différencient sa classe des autres classes ou le différencient des autres signifiés au sein de sa propre classe sont nommés *sèmes* et sont désignés par des paraphrases de longueur quelconque. De la seconde, elle utilise le principe de la détermination du local par le global qui stipule que le sens des unités linguistiques est déterminé par la globalité du texte et son contexte de production et d'interprétation.

La SD est donc une sémantique non compositionnelle qui définit la signification comme un rapport linguistique entre signifiés et qui donne au sens (contenu du mot en contexte) une place prépondérante par rapport à la signification, considérée uniquement comme un type artefact constitué par un linguiste à partir de sens observés dans les occurrences. Cette sémantique est donc profondément ancrée dans les textes qui constituent son objet empirique.

La SD, également appelée sémantique interprétative, ne vise pas la compréhension, activité d'un sujet réel, mais se limite à l'interprétation des énoncés, c'est-à-dire au traitement de la contribution du matériau linguistique à leur sens, sans chercher, par exemple, à rendre compte d'inférences pragmatiques. C'est une sémantique unifiée qui utilise des concepts et principes communs aux trois paliers de description linguistique : celui du mot, pris en compte par la microsémantique, celui de la phrase traité par la mésosémantique et celui du texte par la macrosémantique. Puisque nous nous intéressons à l'apprentissage de lexiques pouvant être utilisés dans des applications TAL, nous nous focalisons essentiellement ici sur la description de la microsémantique, n'abordant que les idées pertinentes pour nous des deux autres niveaux.

La microsémantique s'intéresse au niveau lexical. Elle vise à associer à un morphème une description de son signifié, appelé *sémème*, qui est un ensemble structuré de sèmes. Les morphèmes se combinant en lexies (unités de signification formées d'un ou plusieurs mots), les sémies, signifiés de ces lexies, sont construites à partir des sémèmes de leurs constituants. Par abus de langage, nous considérons dans la suite le sémème comme représentation d'un signifié d'une lexie[15] ou d'un morphème.

Le sème, comme évoqué plus haut, est une relation binaire entre sémèmes marquant soit une différence entre sémèmes sémantiquement proches, soit le partage d'un élément de signification. Un sème ne justifie donc son existence qu'au regard de deux sémèmes entre lesquels il exprime une relation. Il traduit un rapport sémantique précis, plus porteur d'information que les relations lexicales classiques

15. Nous confondrons d'ailleurs parfois également la notion de lexie et de mot.

(synonymie...), et c'est là ce qui fait la richesse d'une formalisation basée sur la SD par rapport à celle adoptée, par exemple, dans une base lexicale telle que WordNet. La SD propose une approche onomasiologique du lexique structuré par des classes de signifiés, où le sens d'un mot est défini par rapport au sens de mots voisins aussi bien sur l'axe syntagmatique que paradigmatique. Les sèmes sont définis comme des relations d'opposition ou d'équivalence au sein de classes de sémèmes. On distingue les sèmes *spécifiques* qui différencient les sémèmes appartenant à une même classe et les sèmes *génériques* qui sont hérités des classes hiérarchiquement supérieures et indexent les sémèmes dans ces classes. L'ensemble des sèmes génériques d'un sémème forme son *classème* et celui de ses sèmes spécifiques son *sémanthème*.

La définition des sèmes est donc liée à des classes sémantiques qui permettent de caractériser la place des sémèmes dans le système de significations qu'ils forment. La plus petite d'entre elles, le *taxème*, regroupe des sémèmes définis différentiellement par leurs sèmes spécifiques et comprenant tous au moins un même sème générique de faible généralité. Ce sème générique, dit *microgénérique*, indexe les sémèmes dans ce taxème. Il permet de définir, à l'échelle du lexique, une relation de l'ordre de l'interchangeabilité des mots dans un contexte donné, les sèmes spécifiques caractérisant, quant à eux, les particularités des sémèmes dans le taxème, mais aussi le structurant. Par exemple 'couteau'[16] et 'cuillère' contiennent le même sème générique /couvert/[17] et sont, par exemple, différenciés par le sème spécifique /pour couper/. Ils appartiennent alors tous deux au taxème noté //couvert//. Le *domaine*, classe de généralité supérieure, est un groupe de taxèmes correspondant à un espace sémantique lié à un type de pratique sociale donné et à l'intérieur duquel il n'existe (en général) pas de polysémie lexicale. Par exemple, les sémèmes 'couteau' et 'cuillère' contiennent un sème dit mésogénérique /alimentation/ dénotant leur appartenance à un domaine. Les *dimensions* sont, quant à elles, des classes de plus grande généralité, en petit nombre, divisant l'univers sémantique en grandes oppositions. 'Couteau' et 'cuillère' contiennent, par exemple, des sèmes dits macrogénériques /concret/ et /inanimé/ notant leur appartenance à des dimensions.

Parmi ces classes, le taxème occupe une place centrale dans la description de systèmes de signification, Rastier notant que c'est d'ailleurs la seule classe nécessaire, tout sémème contenant au moins un sème générique l'indexant dans son taxème de définition. Et c'est à l'intérieur de ce taxème que sont mis en évidence les sèmes spécifiques.

Le fait qu'un élément lexical apparaisse dans un contexte – on se limite en SD au contexte linguistique – a une influence forte sur le contenu de son sémème. Certains sèmes peuvent être inhibés, activés ou propagés. L'effet peut être étudié tant au niveau microsémantique lors de combinaisons de sémèmes de morphèmes en sémies, qu'au niveau mésosémantique qui se focalise sur l'espace allant du syntagme à fonction syntaxique à la phrase et ses connexions immédiates[18], ou au niveau macrosémantique traitant des textes. Nous ne parlons pas ici de ces aspects dynamiques de l'interprétation, le lecteur intéressé pouvant consulter à ce sujet les textes

16. Notation standard d'un sémème.
17. Notation standard d'un sème.
18. La notion de *période* est donc préférée à celle de phrase pour parler de ce palier.

de Rastier. Nous nous contentons d'une vision plus statique des phénomènes qui régissent et résultent de ce placement en contexte. Le point clé est la notion d'*isotopie*, c'est-à-dire de récurrence de sèmes dans les sémèmes des membres d'un énoncé. Pour revenir une dernière fois à l'aspect dynamique du phénomène et en reprenant les termes de Rastier, la présomption d'isotopie permet d'actualiser des sèmes voire *les* sèmes lors de combinaisons de sémèmes. Cette notion conduit d'ailleurs à réduire la polysémie lexicale des constituants d'une chaîne pour ne retenir que les sémèmes contextuellement pertinents. Dans une optique plus statique, les isotopies jouent un rôle fondamental dans l'interprétation, qui consiste d'ailleurs à les répertorier, et ont de l'importance dans la mise en évidence de la cohésion de l'énoncé étudié. Les isotopies[19] spécifiques portant sur le partage de sèmes spécifiques – on parle alors de *molécules sémiques* – sont responsables de l'effet de cohérence textuelle. Les isotopies génériques taxémiques, portant sur des sèmes microgénériques, donnent une impression référentielle locale. Les isotopies génériques domaniales, portant sur des sèmes mésogénériques, donnent une impression référentielle globale. Les isotopies génériques dimensionnelles, portant sur des sèmes macrogénériques, sont responsables des tons (niveaux de langue) et de points de vue globaux. Les isotopies spécifiques dénotent le sujet précis (*focus* en anglais) du texte ou segment étudié (on parle parfois aussi de thème spécifique dénoté par une molécule sémique). Les isotopies génériques, et en particulier les isotopies domaniales, déterminent quant à elles son thème (*topic* en anglais). Par exemple, l'occurrence dans un même texte des lexies *soldat*, *char*, *offensive* et *général* sera révélatrice d'une thématique guerrière, ceci étant attesté par le fait que tous les sémèmes de ces mots soient porteurs du sème générique /guerre/.

2.2.2 Motivations

À l'issue de cette présentation partielle de la SD, nous allons donner certains arguments motivant l'intérêt que nous portons à cette théorie et notre choix de développer une méthodologie d'acquisition en corpus de lexiques sémantiques construits en se basant sur ses principes.

1. La SD est fortement ancrée dans les textes. Elle donne la primauté au sens par rapport à la signification. La microsémantique n'a pas pour but de répertorier tous les sèmes qui organisent la langue en système mais seulement ceux qui résultent de l'interprétation d'un texte ou d'énoncés en contexte. Dans [RCA94], les auteurs montrent comment il est possible de construire manuellement les bases d'une représentation lexicale de la signification par la simple observation de corpus, en se focalisant sur les contextes proches des lexies choisies. C'est cette observation qui permet de constituer les taxèmes valides dans le domaine étudié (la médecine) et d'expliciter les sèmes pertinents. De la même façon, c'est en explorant manuellement des dialogues réels proférés lors de la mise au point d'un document pour utilisateurs d'un logiciel que Beust

19. On parle essentiellement dans cette section d'isotopies sémantiques, laissant de côté les isosémies à fonction syntaxique telles que les accords en genre et nombre.

[Beu98] extrait les éléments initiaux (sèmes...) permettant d'amorcer son système d'interprétation. De telles démarches systématiques de construction de représentations lexicales incitent à penser qu'il est possible de les reproduire de manière automatique. C'est ce que nous montrons au chapitre suivant.

2. La théorie permet de faire émerger une méthodologie pour construire automatiquement des lexiques à partir d'un corpus. Nous avons en effet pu mettre au jour trois étapes pour ce faire, étapes que nous exposons ici en débutant par une synthèse des principes clés de la SD qui nous ont guidés.

   La SD affirme que le domaine est le niveau de structuration de l'espace sémantique au sein duquel peut être associée une interprétation stable à un signe. C'est donc dans un domaine donné qu'il convient de faire émerger les taxèmes, et à l'intérieur de ces taxèmes qu'il faut faire apparaître les sèmes spécifiques permettant de structurer ces classes et de distinguer leurs membres. Un domaine peut être assimilable à un thème générique et donc être mis en évidence par la reconnaissance d'une isotopie, c'est-à-dire par la récurrence de sèmes dans les sémèmes de lexies d'une unité textuelle. Ne disposant pas *a priori* des sémèmes des lexies, nous nous attelons à mettre au jour les lexies dont la cooccurrence est révélatrice du thème. Ces lexies « porteuses » de thèmes nous permettent de scinder un corpus initial non spécialisé en sous-corpus thématiquement homogènes au sein desquels des lexiques peuvent être construits. Un paradigme est défini, en SD, comme l'ensemble des unités pouvant occuper une même place dans un syntagme. Cette hypothèse, qui rejoint tout à fait celles émises par Harris [HGR+89], conduit à faire émerger des taxèmes en étudiant la similarité des contextes dans lesquels apparaissent des lexies. Nous employons pour ce faire une méthode de classification hiérarchique. Enfin, concernant la structuration interne d'un taxème par des sèmes spécifiques, nous proposons d'étudier de manière précise les contextes partagés ou non par les membres de ce taxème, en tentant de déterminer des moyens pour automatiser le plus possible cette tâche.

   Nous définissons donc, en nous basant sur la SD et sans connaissances *a priori*, une méthodologie d'acquisition automatique de relations lexicales intracatégorielles, comportant trois étapes, qui, à partir d'un corpus abordant des thèmes variés – ce qui nous démarque par exemple de [Ass98] qui se place d'emblée dans un domaine donné – permet de construire des lexiques sémantiques pour chacun des thèmes abordés. Nous verrons, au chapitre 3, que si notre recherche sur l'apprentissage de lexiques basés sur la SD est encore un travail en cours, nous faisons des propositions concrètes permettant, à partir d'un corpus, d'aboutir à une représentation sémique. Ceci conduit à montrer que cette théorie autorise *effectivement* à bâtir de tels lexiques par l'observation des mots en corpus.

3. Les relations lexicales sémantiques proposées par la SD sont plus riches que les relations traditionnelles. Les taxèmes regroupent certes des mots « similaires » au sens des études basées sur la linguistique harrissienne qui visent à la constitution de classes sémantiques, dont nous avons cité quelques exemples en section 2.1.2. Cependant la structuration interne de ces taxèmes et l'ana-

lyse sémique de manière générale font émerger entre les mots des relations très fines permettant de mieux saisir leurs rapports, mais aussi d'affiner la description de chaque lexie. L'examen de la représentation du signifié d'une lexie obtenue dans un domaine donné et de la diversité des relations intracatégorielles que cette lexie entretient au sein de plusieurs thèmes conduit à faire apparaître différentes facettes de sa sémantique. On est alors assez proche du niveau d'analyse réalisé manuellement dans [FHL97]. Dans une optique de traitement de variantes sémantiques, cette finesse de relations peut permettre d'envisager l'accès à la variation de sens induite par la reformulation, par une étude des sèmes distinguant les mots échangés.

4. Sur un plan applicatif, la méthode d'acquisition de lexiques sémantiques, mise au point à l'aide de principes de la SD, conduit à obtenir des relations intracatégorielles utilisables dans des applications d'accès au contenu de documents, par exemple pour étendre les requêtes dans un SRI. En plus de ce que l'emploi de relations de synonymie, antonymie... détectables au sein de taxèmes peut apporter en ce sens, l'exploitation de relations sémiques peut permettre de nuancer et contrôler ces expansions de requêtes. La première phase de cette méthodologie, consistant à bâtir des listes de mots porteuses d'un thème, peut aussi être exploitée dans ce type d'applications.

2.3 Le lexique génératif

Nous exposons ici certains principes du Lexique génératif (LG) qui nous sert de cadre théorique pour délimiter des liens nomino-verbaux (N-V) à apprendre en corpus et à exploiter en RI. Nous ne présentons que les points nécessaires à la compréhension de nos travaux explicités au chapitre 4, [Pus95] et [BB01] donnant une vue complète de ce modèle. Nous expliquons également le choix du LG pour l'acquisition de liens N-V.

2.3.1 Description

Le modèle du LG est fondé sur une vision compositionnelle du sens des unités lexicales. Pustejovsky en motive la création par une double limitation des lexiques consistant à énumérer *a priori* tous les sens possibles d'un mot. D'une part, il n'est pas toujours possible de déterminer quel sens est instancié en contexte ; d'autre part, une telle approche ne tient pas compte de la dimension créative de l'utilisation des mots en contexte, le potentiel sémantique d'un mot s'enrichissant au contact de son environnement, ce qui condamne toute tentative de recensement de l'intégralité de ses sens.

Le LG propose une méthode originale en sémantique lexicale pour pallier ces lacunes, qui repose sur les trois postulats suivants :

1. le lexique encode dans les entrées lexicales les propriétés sémantiques nécessaires pour expliquer le comportement linguistique des mots ;

2. il ne s'agit pas d'ensembles non structurés de propriétés syntaxiques ou de types sémantiques. Au contraire, ces informations sont décrites de manière cohérente et structurée dans des représentations lexicales complexes (la *structure des qualia*). Celles-ci définissent la structure interne du mot, c'est-à-dire les différents prédicats indispensables à sa compréhension et la manière de les projeter au niveau syntaxique ;
3. ces représentations lexicales sont manipulées par des *opérations génératives* qui, appliquées aux représentations de base, sont responsables du sens du mot en contexte.

Nous abordons successivement les deux composantes de calcul du sens dans LG, à savoir la représentation lexicale structurée et les mécanismes génératifs d'interprétation des mots en contexte.

Les représentations lexicales en LG consistent en des ensembles structurés de prédicats qui définissent le mot. Comme ces prédicats sont typés, elles peuvent aussi être considérées comme des réserves de types sur lesquelles viennent opérer différentes stratégies interprétatives responsables du sens en contexte. Leur description implique trois niveaux de représentation orthogonaux : les structures argumentale (`argstr`), événementielle (`eventstr`) et des qualia (`qs`), illustrées en figure 2.1 pour un mot quelconque **M** :

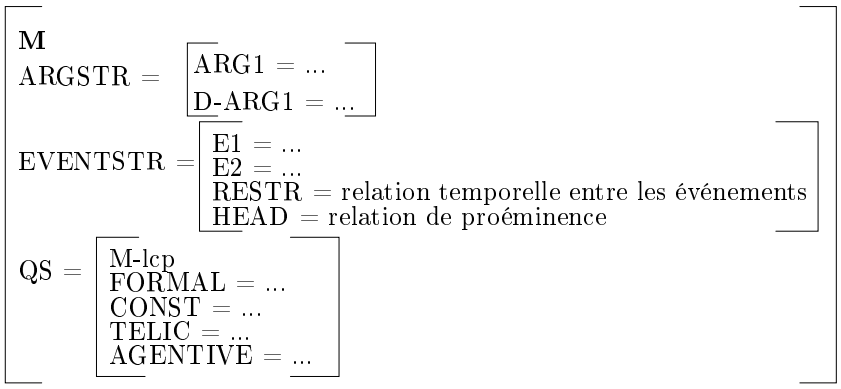

FIG. 2.1 – *Entrée lexicale dans le LG*

Les mêmes niveaux de description sont utilisés pour toutes les catégories syntaxiques. D'une part, les structures argumentale et événementielle définissent les arguments et événements qui interviennent dans la définition des mots. Ceux-ci peuvent être obligatoires ou facultatifs – ils sont dans ce cas appelés *default arguments* (D-ARG) ou *default events* (D-E). D'autre part, la structure des qualia lie ces arguments et événements entre eux et définit leur rôle dans la sémantique du mot.

Dans la structure des qualia, les rôles correspondent à des traits interprétés, qui fournissent le vocabulaire de base pour la description lexicale et déterminent la structure des informations associées à un item lexical donné (c'est-à-dire son

paradigme lexical conceptuel, *lcp*). Le rôle *formel* (FORMAL) encode la fonction d'identité. Celle-ci associe à l'entité sa classe sémantique, par exemple ARTEFACT pour *couteau*, ou le produit des deux types INFORMATION et OBJET_PHYSIQUE pour *livre* (noté INFORMATION.OBJET_PHYSIQUE) ; ces deux types sont liés ici par la relation de contenance, puisque l'objet physique *contient* les informations. Le rôle *constitutif* (CONST) définit les parties de l'objet, comme *pages* ou *couverture* pour le *livre*. Il permet ainsi de représenter adéquatement des mots comme *groupe* ou *partie*. Le rôle *télique* (TELIC) reprend la fonction ou le but de l'objet. Celui-ci est interprété comme un opérateur modal et l'existence du prédicat qui y est encodé ne dépend pas de celle de l'objet. Par exemple, un *livre* peut *être lu*. Inversement, le rôle *agentif* (AGENTIVE) définit la cause ou le mode de création de l'objet : il est interprété comme un quantificateur existentiel, puisqu'il constitue en quelque sorte la condition nécessaire à toutes les autres propriétés de l'objet. Le *livre* ne peut en effet pas *être lu* s'il n'a pas été *écrit*. *Livre* se voit ainsi associer la représentation de la figure 2.2, que nous reprenons de Pustejovsky [Pus95].

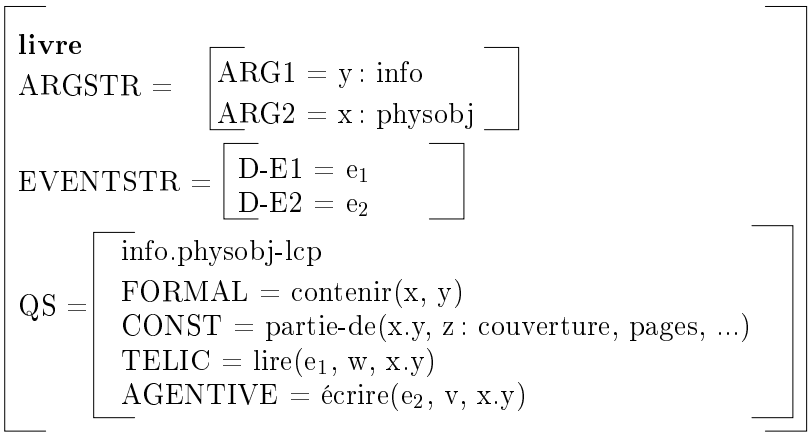

FIG. 2.2 – *Représentation lexicale de livre*

Celle-ci reçoit l'interprétation logique suivante :
$\lambda x.y[livre(x : \texttt{physobj}.y : \texttt{info}) \wedge contenir(x,y) \wedge \lambda w \lambda e_1[lire(e_1,w,x.y)]$
$\wedge \exists e_2 \exists v[écrire(e_2,v,x.y)]]$

Parmi les opérations génératives responsables de l'interprétation des mots en contexte, on distingue :

- *la coercion de type*, qui intervient lorsque l'interprétation sémantique d'un mot est contrainte sous l'influence d'un item qui le gouverne, sans que son type syntaxique soit modifié ;
- *le liage sélectif*, qui intervient lorsqu'un item lexical agit spécifiquement sur une sous-structure d'un syntagme, sans en changer le type ;

– *la co-composition*, qui intervient lorsque plusieurs éléments agissent comme foncteurs et génèrent de nouveaux sens non-lexicalisés.

En contexte, ces opérations génératives contrôlent la projection des expressions relationnelles définies dans la structure des qualia, en déterminant leur combinaison avec celles des mots voisins. L'information prédicative associée à un N doit, en particulier, être accessible pour rendre compte de façon satisfaisante de certains mécanismes linguistiques. Prenons un exemple emprunté à [Pus95] et traduit pour illustrer la première opération. Pustejovsky explique l'interprétation de *Jean commence un livre* par le fait que *commencer*, qui requiert un objet de type événementiel, impose à *livre* un changement de type, d'objet physique à événement. Le nom *livre* projette le type sémantique requis par la règle de coercion[20], soit l'information relationnelle télique (*Jean commence à lire un livre*) soit l'information agentive (*Jean commence à écrire un livre*).

Un des objectifs de Pustejovsky est de montrer que les structures proposées (manipulées par les opérations génératives) permettent de représenter adéquatement la *polysémie logique* des expressions linguistiques, entre autres les alternances verbales (1) et nominales (2), les différences dans la forme syntaxique d'un argument (3) et les changements aspectuels (4).

1. *Alternances verbales*
   a. je commence la symphonie (TRANSITIF)
   b. la symphonie commence (INTRANSITIF)
2. *Alternances nominales*
   a. je prends mon repas avec moi (NOURRITURE)
   b. pendant le repas, j'ai dormi (ÉVÉNEMENT)
3. *Différences dans la forme syntaxique*
   a. je commence le livre (+SN)
   b. je commence à lire le livre (+SV)
4. *Différences aspectuelles*
   a. « *I bake a cake* » (je fais cuire un gâteau) (ACCOMPLISSEMENT)
   b. « *I bake potatoes* » (je fais cuire des pommes de terre) (PROCÈS)

Pour chaque mot, et en particulier pour chaque N – cette catégorie constituant notre objet d'étude –, le LG permet de définir un réseau de relations qui lui sont associées lexicalement et qui ont leur propre interprétation, par exemple *livre lire*, *livre écrire* et *livre contenir* pour *livre*. Celles-ci ne sont pas définies empiriquement, mais sont motivées linguistiquement : il s'agit des relations *nécessaires* pour expliquer le comportement sémantique du mot. Par exemple, c'est parce que le *livre* est un contenant qu'il est possible de parler d'un *livre d'images*. De même, c'est parce que sa sémantique fait référence à la fonction et au mode de création, que l'on peut dire *je commence un livre* dans le sens de *je commence à le lire* ou à *l'écrire*.

20. Si Godard et Jayez [GJ93] discutent ce principe de coercion du type d'un argument, ils ne remettent pas en cause celui de projection d'une information prédicative à partir du nom.

2.3.2 Motivations

Nous explicitons ici quelques arguments justifiant notre intérêt pour le formalisme du LG et pour l'apprentissage de liens nomino-verbaux dans lesquels le V instancie l'un des rôles de la structure des qualia du N (nous parlons par la suite de lien, paire ou couple qualia à ce sujet).

1. Le LG est un cadre formel qui, s'il ne théorise pas, dans ses grands principes, la dimension textuelle, reste compatible avec un ancrage dans les textes et partage donc cette caractéristique avec la sémantique différentielle. S'il associe théoriquement par défaut des représentations aux mots dans le lexique, celles-ci sont sous-spécifiées et vont être enrichies, voire modifiées par le contexte. Un *livre*, par exemple, pourrait aussi être *publié*, *imprimé*, ou même *rangé*. Il peut aussi servir à *enseigner* ou à d'autres tâches. Il est donc à la fois nécessaire et justifié d'apprendre en corpus les prédicats téliques, agentifs... de noms dont on étudie la sémantique.
2. Les liens N-V qualia sont intéressants pour caractériser des variantes sémantiques de N et désambiguïser ces noms, et ils sont exploitables en RI. Nous avons déjà illustré en section 2.1.3 la force et l'intérêt du lien N-V tant pour générer ou reconnaître des variantes de termes complexes [Fab96] que pour préciser le sens des N [Gre97]. Or, pour chaque N, le LG instaure des relations prédicatives nécessaires pour expliquer son comportement sémantique[21]. Si l'on accepte donc que le LG définit les propriétés lexicales intéressantes d'un point de vue sémantique et que ces propriétés sont instanciées en contexte, il devient aussi possible d'utiliser les structures des qualia pour organiser, ou structurer, les informations contenues dans le texte. Les relations exprimées dans les structures des qualia sont ainsi des données lexicales privilégiées pour la recherche d'information. Nous systématisons donc la proposition de [Gre97] d'utiliser des paires N-V en RI et définissons une paire N-V comme pertinente si elle est qualia, c'est-à-dire si le verbe instancie l'un des rôles de la structure des qualia du N. Différentes propositions étayent d'ailleurs déjà cette hypothèse. Cécile Fabre [Fab96] a montré que les liens N-V exprimés dans les qualia permettent de calculer la représentation sémantique des groupes nominaux et nous avons proposé d'utiliser ces liens pour étendre une requête [FS99]. Les structures des qualia peuvent aussi servir à alimenter une *toile lexicale* (*lexical web* selon [PBV+97]), c'est-à-dire un réseau de termes pertinents et de relations qui, ensemble, définissent le sujet d'un texte, à la manière d'un index traditionnel. Ce réseau, mis à plat et présenté comme l'index d'un livre, permet à l'utilisateur de naviguer dans le texte. Pour *disk*, par exemple, [PBV+97] propose l'ensemble des relations N-V suivantes qui, selon les auteurs, définissent extensionnellement le sens du mot dans le domaine traité (la

21. Même si Kilgarriff [Kil01] considère que les structures des qualia et les opérations génératives dans le LG ne sont absolument pas suffisantes pour expliquer l'intégralité du potentiel créatif d'un N, ce formalisme ne résistant pas, pour lui, à l'épreuve du corpus, il ne met toutefois pas en cause la nécessité des relations décrites dans ces structures.

documentation technique de Macintosh, *Macintosh Reference*) :

```
disk access
     erase
     name
     initialize
     test
     save file on
     repair
```

Nous souhaitons donc apprendre en corpus ces liens N-V qualia et tester effectivement leur apport dans un SRI.

3. L'apprentissage par une méthode « explicative » de paires N-V qualia permet de contribuer à la réflexion linguistique sur la définition des différents rôles qualia. Le LG est un modèle de sémantique lexicale neuf, qui s'affine encore actuellement (voir par exemple [BCL01] sur la notion de qualia étendue). Apprendre des liens N-V qualia par une méthode d'apprentissage produisant des règles explicatives, comme la programmation logique inductive (*cf.* chapitre 4), permet alors de faire émerger certains fondements linguistiques de cette notion de rôles qualia, comme, par exemple, la connaissance des structures portant un rôle donné.

2.4 Conclusion

Dans ce chapitre, nous avons explicité, en illustrant leur nécessité par un besoin applicatif, les éléments que nous voulons acquérir en corpus pour enrichir la description lexicale de noms dans une double optique de désambiguïsation et de traitement de variations sémantiques, ainsi que les cadres théoriques dans lesquels nous nous situons pour ce faire. Ce placement dans des cadres linguistiques est fondamental, pour nous permettre de valider ce que nous apprenons et pour nous guider dans la mise au point de méthodes d'apprentissage.

Nous nous intéressons, via un placement dans la SD, à l'acquisition de relations sémantiques intracatégorielles « traditionnelles » d'une part, c'est-à-dire de liens synonymiques, antonymiques..., mais aussi à celles de liens sémantiques plus fins, dénotant plus précisément les diverses facettes sémantiques d'un nom et les relations qu'il entretient avec les membres de son paradigme. Nous visons également, via le placement dans le LG, à enrichir la description lexicale de N de liens N-V permettant d'expliquer des mécanismes inférentiels, mécanismes d'interprétation fondamentaux tels que la coercion de type.

Avant de passer à l'exposé de ces travaux, nous soulevons toutefois la question de l'articulation entre les deux cadres théoriques que nous utilisons ; en d'autres termes, aurait-il été possible, pour répondre à l'objectif d'acquisition d'éléments permettant d'enrichir la description des noms que nous nous sommes fixé, de nous

limiter à un seul de ces deux formalismes? La réponse est clairement négative[22]. Les potentialités du LG à expliciter des liens N-N pouvant servir à des reformulations de noms sont très limitées. Les seuls liens (éventuellement) intracatégoriels pris en compte dans le LG sont des liens hyperonymiques (rôle formel) et méronymiques (rôle constitutif). Le LG est un formalisme qui se focalise sur quatre aspects du sens des mots (les rôles qualia) et ne cherche pas à établir des relations de synonymie, antonymie... également utiles pour la reformulation. La SD, quant à elle, organise les mots au sein de classes paradigmatiques monocatégorielles. Dans cette théorie, tout élément différentiel est potentiellement définitoire ; pour reprendre un exemple classique de ce formalisme, il est possible, par exemple, de caractériser le taxème des sièges par un sème générique /pour s'asseoir/, qui laisse éventuellement apparaître un prédicat verbal associable à tous les noms du taxème. Cependant le nommage des sèmes est peu systématisable[23] et leur nom est simplement une paraphrase linguistique quelconque reflétant le type de nuance de sens détectée. On ne peut donc envisager d'exploiter ce cadre formel pour accéder de manière simple et automatique à des liens intercatégoriels, alors que, de son côté, le LG met en avant des prédicats verbaux intéressants pour caractériser des N. Pour nous, chaque formalisme complète l'autre en permettant d'accéder à différents éléments de la sémantique des N utiles pour la prise en compte de la variation sémantique et la désambiguïsation.

Nous débutons par l'exposé de nos réalisations concernant la mise au point d'une méthodologie complète d'apprentissage en corpus de lexiques sémantiques basés sur les principes de la SD.

22. Dans la version actuelle de ces deux modèles ; nous n'avons pas cherché à étudier des possibilités éventuelles d'extensions de ceux-ci.

23. Nous verrons, au chapitre suivant, que ce qui est réalisable est la mise en évidence de l'existence d'un sème générique (pour un taxème) ou spécifique (au sein d'une telle classe), et la caractérisation de celui-ci par des éléments du contexte des mots impliqués qui peuvent aider à comprendre et à nommer ce sème.

Chapitre 3

Apprentissage de relations intracatégorielles basées sur la sémantique différentielle

Dans ce chapitre, nous présentons une synthèse de recherches que nous avons menées et dont nous poursuivons le développement, qui ont pour objectif de mettre au point une méthodologie d'acquisition en corpus, sans connaissances *a priori*, de lexiques sémantiques basés sur la sémantique différentielle de Rastier.

Si l'intérêt de l'utilisation des principes et de représentations lexicales issus de cette théorie a déjà été démontré par plusieurs travaux, que ce soit pour l'interprétation de dialogues ou des textes [Beu98, Tan97, RCA94], la diffusion ciblée de documents [Pin99] ou la création d'ontologies pour la consultation de documentations techniques [Ass98] par exemple, l'originalité de notre recherche est d'étudier le développement d'une méthodologie complète de construction de telles représentations, en partant d'un corpus abordant un nombre quelconque de thèmes et sans autre information que l'étiquetage morpho-syntaxique de celui-ci, et en visant la détermination de relations sémantiques intracatégorielles « classiques » (synonymie...) mais surtout plus spécifiques (relations sémiques). Nous essayons également, en travaillant en particulier sur les méthodes statistiques d'apprentissage, d'automatiser le plus possible les différentes phases de l'acquisition. Nous nous éloignons donc en cela aussi des travaux précédemment cités qui construisent ou initialisent manuellement leurs représentations lexicales ou se placent dans un domaine particulier.

Ce chapitre est formé de deux parties. Dans la section 3.1, nous présentons une première version de notre méthodologie, telle que décrite dans [PS00, PS99]. Comme nous l'avons évoqué en section 2.2.2, celle-ci se décompose, en respectant les principes de la SD, en une phase de segmentation du corpus initial en sous-corpus thématiquement homogènes, puis, au sein de chaque sous-corpus, en la création de classes sémantiques taxémiques et l'exploitation des contextes linguistiques des divers éléments de ces classes pour mettre au jour des sèmes spécifiques. Nous abor-

dons également à ce niveau l'exploration des sémèmes d'un même mot à travers différents thèmes pour pointer diverses facettes de sa signification. La méthodologie, telle que décrite dans cette section, requiert cependant des interventions humaines et est perfectible. La section 3.2 présente nos développements en ce sens. Nous montrons en particulier le travail que nous avons réalisé sur l'amélioration de l'adéquation de la méthode de classification utilisée pour construire des listes de mots caractéristiques des thèmes présents dans le corpus initial. Ceci permet d'une part d'accroître la qualité de détection des thèmes et, d'autre part, de ne plus requérir d'expert pour choisir dans l'arbre de classification hiérarchique les classes pertinentes. Nous débutons seulement actuellement le perfectionnement des phases suivantes (construction de taxèmes et détection de sèmes). Sur ce point, nous présentons donc uniquement certaines pistes que nous allons explorer et replaçons nos travaux dans le contexte d'autres recherches qui s'intéressent de manière plus ou moins automatique à l'exploitation du contexte syntagmatique des mots au sein de classes sémantiques. Nous discutons également de la mise en évidence de relations lexicales traditionnelles à partir des méthodes de caractérisation de sèmes, et soulevons la question de la variation de sens induite par le remplacement, dans un cadre applicatif par exemple, d'un mot par une lexie à laquelle il est lié par des sèmes spécifiques, dont l'étude est aussi une de nos perspectives.

Le choix d'une présentation chronologique de nos travaux s'explique par deux raisons. La première, d'ordre pratique, provient du fait qu'un exposé complet des recherches réalisées sur la caractérisation des thèmes, suivi de celui des études concernant les autres phases, nous auraient conduite à mener le lecteur dans une évolution de conditions d'expérimentation un peu difficile à suivre. La seconde, plus importante à nos yeux, est que le séquencement choisi permet de mettre en évidence un des aspects de notre façon d'aborder le TAL (*cf.* chapitre 1), qui consiste à travailler sur les algorithmes d'apprentissage en tant que tels.

Au cours de cette introduction, nous avons, à plusieurs reprises, indiqué notre souci d'automatiser le plus possible les diverses phases de notre travail, en particulier celles concernant la formation des classes de lexies porteuses de thèmes et des taxèmes. Or, certains auteurs, dont Bourigault [Bou02, AB96] ou Habert [NZHB01, FHL97] par exemple, partent plutôt du principe que, pour la constitution et la validation de classes, une interaction importante avec des experts est nécessaire, avec retour au contexte. Plus que d'outils « opaques » produisant un résultat donné, ils cherchent davantage à obtenir des produits semi-finis, dont les propositions sont modifiables. Nous nous plaçons, pour notre part, dans une optique différente de ces travaux : notre objectif est de proposer une méthodologie efficace et répétable pour produire, pour des applications, des lexiques sémantiques adaptés à leurs domaines. Les classes, tant celles révélatrices de thèmes que les classes sémantiques, font partie d'une chaîne de traitement allant du corpus à une organisation sémique des taxèmes, que nous cherchons à rendre autonome, d'où notre souci de mettre au point des algorithmes performants d'apprentissage de ces classes.

Les recherches qui sont présentées ici ont été développées en collaboration avec des membres de l'Irisa, et leur doivent donc beaucoup : Ronan Pichon, Israël-César Lerman qui nous apporte son soutien sur les aspects statistiques, et Mathias Rossi-

gnol qui a débuté en octobre 2001 une thèse sous mon encadrement.

3.1 Méthodologie d'acquisition

L'objectif de notre méthodologie d'acquisition de relations intracatégorielles basées sur la SD est, à partir d'un corpus étiqueté morpho-syntaxiquement, d'aboutir, au sein de taxèmes appris, à la mise en évidence de sèmes spécifiques entre les lexies regroupées. Nous avons montré en section 2.2.2 comment la théorie de Rastier nous offrait des guides pour définir trois étapes successives pour ce faire. Pour plus de clarté, nous en rappelons ici les principes, ce qui nous permet également d'expliciter la démarche que nous avons suivie à ces différents paliers.

La SD considère le domaine comme le niveau de structuration de l'espace sémantique au sein duquel il est possible d'assigner une interprétation stable à un signe. Il est assimilable à un thème générique et peut donc être mis en évidence par la reconnaissance d'une récurrence de sèmes dans les lexies d'un segment textuel. Ne disposant pas *a priori* des sémèmes des mots pour reconnaître ces isotopies (puisque nous voulons les construire), nous avons choisi de mettre au jour les lexies dont la cooccurrence est révélatrice d'un thème en constituant ces ensembles de mots par une classification hiérarchique fondée sur la similarité des distributions des mots dans les différents paragraphes du corpus. Ces listes de lexies « symptomatiques » de thèmes sont utilisées pour scinder un corpus initial non spécialisé en sous-corpus thématiquement homogènes, en utilisant la coprésence d'un certain nombre de leurs membres dans un segment de texte pour affecter un thème à celui-ci.

C'est au sein de ces sous-corpus qu'il est possible de construire des lexiques, et donc de faire émerger des taxèmes à l'intérieur desquels on peut distinguer les éléments par des sèmes spécifiques. Un paradigme étant défini en SD comme l'ensemble des unités pouvant occuper une même place dans un syntagme, nous utilisons cette hypothèse pour faire apparaître, par classification, des taxèmes de lexies en nous basant sur la similarité de leurs contextes.

Concernant la troisième phase de structuration des classes sémantiques par des sèmes spécifiques, nous proposons d'étudier les contextes partagés ou non par les membres d'un taxème, en déterminant des moyens pour automatiser le plus possible cette tâche.

Dans cette section, nous présentons et discutons les résultats obtenus par la mise en place de cette méthodologie sur le corpus du *Monde diplomatique* (7,8 millions de mots provenant d'une sélection d'articles datant de 1987 à 1997 parmi ces archives du mensuel) étiqueté à l'aide des outils Multext (Mtseg et MtLex) de l'Université de Provence [IV94] et désambiguïsé par le logiciel Tatoo de l'Issco de Genève [ABR95], en reprenant des données décrites en détail dans [PS00, PS99]. Ce corpus présente l'avantage pour nous d'aborder une grande variété de thèmes (géopolitique, macroéconomie, art, sociologie...). Nous débutons par la phase de détermination des listes de mots caractérisant les divers thèmes, présentons ensuite le travail effectué sur la production de taxèmes et leur structuration par des sèmes, puis tirons un bilan de cette première version de notre méthodologie.

3.1.1 Caractérisation des thèmes

Comme nous venons de le voir, l'idée clé de notre technique de détermination des thèmes d'un corpus est de chercher à détecter les isotopies sémantiques caractéristiques de ces thèmes en étudiant leur manifestation dans les paragraphes de ce corpus à l'aide de la cooccurrence de certains mots au sein de ces unités textuelles. Nous n'avons aucun *a priori* sur le nombre de thèmes abordés dans le corpus, ni sur le contenu ou sur la taille des listes de mots symptomatiques de ces thèmes. Toutefois, les mots porteurs d'un thème apparaissant fréquemment dans les paragraphes abordant ce thème, ils doivent donc posséder des répartitions similaires sur l'ensemble des paragraphes du corpus. C'est ce raisonnement qui nous a conduite à la mise au point d'une méthode de classification hiérarchique fondée sur la distribution relative des noms dans les paragraphes du corpus pour détecter et caractériser les thèmes.

Détection de thèmes : comparaison à l'existant

Plusieurs travaux se sont intéressés à la détection automatique de thèmes en s'appuyant sur des indices linguistiques [LP95] ou sur des notions telles que la cohésion lexicale [FG01, Hea94], certains d'entre eux réalisant simultanément la caractérisation de thèmes et la segmentation du discours. Les dernières recherches citées étant plus proches de nos préoccupations, nous nous attardons ici quelque peu sur elles pour mieux dissocier nos méthodes et objectifs.

TextTiling [Hea94] est un outil de segmentation d'un texte en groupes de paragraphes successifs portant sur le même thème, qui se base sur une mesure de similarité lexicale entre séquences consécutives de mots. Tous les 20 mots environ, l'algorithme calcule la ressemblance entre les listes de 100 mots apparaissant à droite et à gauche du point de focus. Un minimum local de cette mesure est alors considéré comme un indice de zone de changement thématique, dont la frontière est ramenée à la limite de paragraphe la plus proche. Les mots ayant eu un rôle proéminent dans le maintien à une valeur élevée de la mesure entre deux minima sont utilisables pour caractériser le thème de la région considérée. Cependant, rien ne garantit que deux zones séparées traitant d'un même thème soient caractérisées par des mots identiques, ce qui rend difficile la détection automatique de proximité thématique entre segments non consécutifs. [FG01] est basé sur une idée similaire mais réalise une segmentation à granularité beaucoup plus fine, la mesure de consistance lexicale étant calculée pour chaque mot et à l'aide de fenêtres de plus ou moins 10 mots autour du point de focus. Puisque le processus manipule moins de données, des informations supplémentaires sont utilisées pour enrichir le calcul : un corpus de 45 millions de mots sert préalablement à extraire un réseau de collocations, et cette connaissance est exploitée dans la mesure de proximité lexicale pour réaliser une première segmentation des textes étudiés. À l'aide de cette segmentation, le système définit des « signatures thématiques » qui forment la base d'une seconde segmentation et fournissent une caractérisation indirecte des thèmes détectés.

Nos objectifs sont différents de ceux de ces travaux : nous ne nous intéressons pas à une structuration de la lecture séquentielle des textes, mais cherchons à déter-

miner une caractérisation permettant de connaître de manière immédiate le thème abordé dans un segment du corpus. Nous ne réalisons donc pas une analyse linéaire de l'intégralité de celui-ci. De plus, nous ne voulons pas avoir recours à des connaissances extérieures, comme c'est le cas dans [FG01]. Nous prédéfinissons le paragraphe comme subdivision de texte consistante au niveau du thème, en nous basant sur le type de corpus que nous manipulons. Notons toutefois que la méthode que nous avons mise au point nous autorise à affecter un paragraphe donné à plusieurs sous-corpus thématiques si nécessaire.

Caractérisation des thèmes par classification hiérarchique

La première expérience d'apprentissage de listes de lexies caractéristiques de thèmes a été réalisée sur un sous-ensemble d'un million de mots du corpus du *Monde diplomatique* (200 articles, 9500 paragraphes). Nous avons sélectionné des noms suffisamment fréquents (165 noms ont été retenus) et avons adjoint à chaque lemme son nombre d'occurrences dans les différents paragraphes du corpus d'apprentissage.

La classification basée sur la distribution relative des lemmes à travers les paragraphes a été effectuée à l'aide de `chavleps` [PLL92], implémentation de la méthode CHAVL de classification hiérarchique par analyse de la vraisemblance des liens développée par I.-C. Lerman [Ler91]. La particularité de cette technique par rapport à d'autres méthodes de classification ascendante hiérarchique réside dans sa façon de choisir les classes à fusionner à un niveau de la classification. L'analyse de la vraisemblance des liens (AVL) prend en considération deux facteurs : d'une part, la cohérence globale de l'ensemble qui serait formé en rassemblant les deux classes étudiées, sorte de mesure de « densité » de cet ensemble ; d'autre part, la vraisemblance de l'existence de ce regroupement en tant que classe qui est évaluée en calculant si sa densité est réellement statistiquement « exceptionnelle » eu égard aux cohérences respectives des deux classes considérées individuellement. C'est cette seconde évaluation qui fait l'originalité et l'efficacité de l'AVL comme mesure de cohérence (et lui donne son nom). Une façon habituelle de lire un arbre de classification consiste à effectuer une coupure de ses branches à un niveau donné et à extraire les classes obtenues. Afin de guider le choix du niveau de lecture, CHAVL donne à chaque étape du calcul une « note » reflétant la cohérence interne des classes de la partition correspondante. Il est ainsi possible de suivre les évolutions de cette mesure de qualité au cours de la constitution de l'arbre et d'en détecter les maxima locaux, susceptibles d'indiquer qu'une « bonne » classification a été atteinte.

Résultats

La lecture à un niveau unique de l'arbre de classification obtenu lors de notre apprentissage n'étant pas adaptée au but visé[1], les 80 classes de 3 à 15 mots présentes dans cet arbre ont été extraites. Elles ont été évaluées par 5 personnes qui, si elles estimaient la classe révélatrice d'un thème, devaient le nommer. 27 classes ont ainsi été jugées pertinentes (accord d'au moins 4 personnes), dont {*journal*, *journaliste*,

1. Des classes potentiellement intéressantes sont en effet présentes à plusieurs niveaux de l'arbre de classification.

presse} révélatrice du thème PRESSE, ou {*international*, *communauté*, *organisation*, *nation*, *développement*} du thème ORGANISATIONS (*cf.* [PS00] pour plus de détails).

Nous avons cherché à voir s'il était possible de limiter de manière automatique le nombre de classes à évaluer. Ainsi, en prenant en compte le critère de qualité de l'AVL et en favorisant les classes formées au niveau des maxima de cette mesure, 45 classes parmi les 80 sont identifiables. Parmi celles-ci, 21 des 27 classes recensées comme reflétant un thème sont présentes.

En utilisant le principe de la coprésence d'au moins deux mots d'une liste caractéristique d'un thème dans un paragraphe pour reconnaître ce dernier comme abordant ce thème, les 27 listes obtenues nous ont conduite à découper en autant de sous-corpus thématiques le corpus initial de 7,8 millions de mots.

3.1.2 Constitution et structuration de taxèmes

Au sein de chaque sous-corpus, l'étape suivante consiste à déterminer des taxèmes en se basant sur la similarité des contextes d'apparition des mots. Nous avons choisi, pour ce faire, d'utiliser une méthode déjà éprouvée[2]. Cette décision s'explique par le fait que l'objectif de cette première version de la méthodologie est de montrer la faisabilité du passage « corpus non spécialisé → lexiques basés sur la SD » en mettant en évidence les points où des efforts seront à effectuer, et que l'originalité de nos recherches, au sein d'un thème, réside, plus particulièrement, en la tentative de structuration interne des taxèmes par des sèmes.

Constitution de taxèmes par classification hiérarchique

Nous avons ainsi associé à chaque N suffisamment fréquent pour qu'une analyse statistique soit fondée, un vecteur des N, V et A (et leur nombre d'occurrences) apparaissant dans des fenêtres de plus ou moins 5 mots autour de ses diverses apparitions dans le sous-corpus. Nous avons ensuite réalisé une classification en utilisant `chavleps`, avec pour mesure de similarité le produit scalaire normalisé entre les vecteurs de contexte. Dans cette version de la méthodologie, les éléments de contexte sont donc vus comme des ensembles non structurés de mots, dont les positions par rapport à la lexie étudiée ne sont pas différenciées. Pour chaque classe et pour chacun des mots la constituant, nous mémorisons les co-textes associés. Nous n'avons pas, à ce niveau, évalué la qualité des classes sémantiques obtenues : nous sélectionnons à la main certaines d'entre elles pour étudier la possibilité d'y faire apparaître une organisation sémique. Nous reviendrons en section 3.2.2 sur le travail à réaliser, en particulier pour accroître la qualité des classes et automatiser leur sélection dans l'arbre de classification.

2. De nombreux travaux, dont on connaît les points forts et les limites, ont déjà été réalisés sur la constitution de classes sémantiques à partir de corpus (*cf.* [Res93, Gre94b, WSG96] par exemple).

Structuration des taxèmes

Dans les taxèmes retenus, nous avons cherché à mettre au jour des blocs de contexte caractérisant un de leurs membres par rapport aux autres, ou spécifiques de la signification d'un mot dans un thème par rapport à sa signification dans un autre thème. Nous présentons dans [PS99] un certain nombre de résultats concernant cette mise en évidence de sèmes spécifiques. Nous en détaillons quelques-uns ici.

Dans ce travail, lors de la constitution des taxèmes, les N et les A apparaissant le plus fréquemment dans le voisinage des N à classer sont mémorisés[3]. Ces voisinages simplifiés nous servent, par calcul d'intersections et de différences ensemblistes, à mettre au jour des différenciations de sens entre mots. Notre but est d'interpréter, actuellement manuellement, les ensembles de mots de contexte obtenus par ces opérations simples, en y recherchant des séquences caractérisant des traits sémantiques particuliers. Les membres de ces séquences ont la particularité de posséder un élément de sens en commun, ce qui conduit à leur désambiguïsation implicite. Les mots de voisinage trop ambigus ou isolés ne sont pas pris en compte, et les différents éléments de sens ainsi mis en évidence sont associés au mot étudié.

Taxème {*pouvoir, autorité, gouvernement*} dans le thème NÉGOCIATIONS
Au sein du thème NÉGOCIATIONS (caractérisé par la liste de mots {*négociation, accord, création, position*}), les voisinages simplifiés des mots de la classe sémantique {*pouvoir, autorité, gouvernement*} sont :

pouvoir : *accession* 8, *an* 8, *armée* 8, *concentration* 8, *pays* 8, *nouveau* 9, *place* 9, *coalition* 10, *contrôle* 10, *gouvernement* 10, *arrivée* 16, *état* 17, *partage* 17, *parti* 17, *achat* 22, *central* 22, *public* 27, *économique* 28, *politique* 50

autorité : *américain* 3, *frontière* 3, *local* 3, *nouveau* 3, *pays* 3, *pouvoir* 3, *problème* 3, *provisoire* 3, *armée* 4, *autonome* 4, *Cisjordanie* 5, *élu* 5, *état* 5, *gouvernement* 5, *politique* 7, *palestinien* 8

gouvernement : *actuel* 11, *opposition* 11, *position* 11, *premier* 11, *sandiniste* 12, *accord* 13, *membre* 13, *national* 13, *central* 14, *chef* 14, *occidental* 14, *état* 16, *Bonn* 17, *coalition* 17, *formation* 18, *français* 25, *américain* 26, *nouveau* 26, *pays* 26, *fédéral* 27, *européen* 28, *politique* 37, *israélien* 61

Les points communs à ces trois voisinages sont {*état*, *nouveau*, *pays*, *politique*} ; la présence des premier et troisième mots montre que l'état ou le pays sont représentés par un pouvoir, une autorité ou un gouvernement. Le fait que *gouvernement* apparaisse dans les contextes d'*autorité* et *pouvoir* laisse apparaître une possibilité de hiérarchie hyperonymique. On note également, au chapitre cette fois des différences, qu'*autorité* semble associé à une notion de précarité (*problème*, *provisoire*, *autonome*), tandis que *gouvernement* dénote un pouvoir institutionnel (*fédéral*, *national*) et structuré (*chef*, *membre*, *coalition*, *formation*), qui représente quelque chose ou quelqu'un (*sandiniste*, *occidental*, *français*, *américain*, *européen*, *israëlien*). *Pouvoir*, enfin, est associé à un domaine de compétence précis (*public*, *économique*, *politique*) et semble plus fluctuant (*accession*, *an*, *arrivée*, *partage*).

3. Nous n'avons volontairement pas retenu les V, car l'absence de différenciation de la position des mots dans le contexte d'un N donné les rend peu exploitables.

Taxème {*pouvoir, autorité, gouvernement*} dans le thème TERRITOIRE
Les voisinages des mêmes mots au sein du sous-corpus thématique TERRITOIRE (caractérisé par la liste {*autorité, région, territoire*}) sont les suivants :

pouvoir : *état* 7, *local* 7, *soviétique* 7, *année* 8, *exécutif* 9, *parti* 9, *prise* 9, *public* 10, *économique* 11, *président* 11, *nouveau* 12, *place* 12, *arrivée* 17, *politique* 21, *central* 36

autorité : *Pékin* 4, *place* 4, *président* 4, *preuve* 4, *région* 4, *transfert* 4, *chinois* 5, *état* 5, *nouveau* 5, *territoire* 5, *gouvernement* 6, *politique* 6, *israélien* 13, *palestinien* 13, *local* 16

gouvernement : *fédéral* 7, *occidental* 7, *président* 8, *français* 9, *ministre* 9, *régional* 9, *union* 9, *formation* 10, *politique* 12, *nouveau* 14, *central* 15, *national* 16, *israélien* 32

Dans ce thème, on retrouve peu d'éléments de voisinage communs aux trois mots : {*nouveau, politique, président*}. Toutefois, les mots désignant l'étendue géographique ou institutionnelle sur laquelle l'*autorité*, le *pouvoir* ou le *gouvernement* exerce son autorité sont très fréquents, sans qu'il s'agisse des mêmes pour chacun de ces trois mots ; ainsi {*local, central*} pour *pouvoir*, {*région, territoire, local*} pour *autorité* et {*fédéral, régional, union, central, national*} pour *gouvernement* forment un ensemble cohérent par rapport à cette notion d'étendue géographique (particulièrement pour *autorité*) ou institutionnelle (particulièrement pour *gouvernement*). Parmi les différences, on peut noter que l'*autorité* est très associée à *local* quand *pouvoir* et *gouvernement* sont fortement liés à *central*, ce qui amène à penser que l'*autorité* est subordonnée à un *gouvernement* ou un *pouvoir*. Par ailleurs, la coprésence spécifique de {*fédéral, national*} d'une part, et {*ministre, union, formation*} d'autre part, indique que le *gouvernement* exerce son autorité dans un cadre institutionnel bien défini et structuré, alors que *pouvoir* et *autorité* impliquent un cadre plus informel. L'*autorité* est très liée à la notion de territoire (*région, territoire, local*), alors que le *pouvoir* s'exerce sur autre chose (*public, économique, exécutif*) et semble indiquer une entité plus changeante (*place, prise, arrivée, année*) que *gouvernement* ou *autorité*.

Représentation lexicale

Ces résultats peuvent être exploités pour faire des propositions concrètes d'implémentation de lexiques sémantiques basés sur la SD, en partant de la seule étude du corpus. Ainsi, on peut se servir des séquences de contexte mises en évidence ci-dessus pour bâtir une représentation partielle des significations de *pouvoir*, *autorité* et *gouvernement* dans les thèmes TERRITOIRE et NÉGOCIATIONS.

La figure 3.1 en donne une vue synthétique sous la forme d'un graphe, dans lequel les nœuds indiquent la signification d'un mot dans un thème, et les arcs orientés entre ces nœuds indiquent en quoi chaque signification diffère d'une autre. De façon plus précise, un arc orienté étiqueté /sème/ entre les nœuds A et B indique que le sème /sème/ participe à la signification de A et pas à celle de B, ce sème étant une proposition d'abstraction de l'élément de sens associé aux séquences extraites

précédemment. Par exemple, l'arc étiqueté par le sème /changeant/ liant *autorité*$_{\text{NÉGOCIATIONS}}$ à *gouvernement*$_{\text{NÉGOCIATIONS}}$ reflète le caractère précaire de l'autorité que nous avons souligné par la présence de la séquence *{problème, provisoire, autonome}* dans le contexte de ce seul mot ; l'arc étiqueté /étendue/ entre *gouvernement*$_{\text{TERRITOIRE}}$ et *gouvernement*$_{\text{NÉGOCIATIONS}}$ souligne l'aspect d'étendue sur laquelle un gouvernement exerce son autorité, pointé dans le thème TERRITOIRE par la séquence *{fédéral, régional, union, central, national}*, mais absent dans le thème NÉGOCIATIONS.

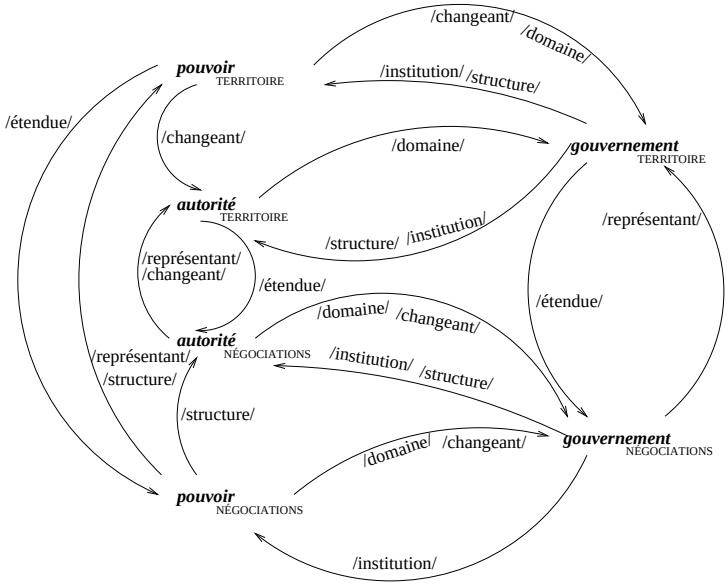

FIG. 3.1 – *Exemple de représentation lexicale*

[PS99] décrit également l'étude d'autres mots comme *militaire*, dont le contexte simplifié permet de faire apparaître une connotation guerrière dans le thème TERRITOIRE (*opération, massif, occupation, victoire, moyen*), par rapport à son utilisation dans le thème NÉGOCIATIONS qui met plus en avant son côté organisé et structuré (*organisation, OTAN, atlantique, dépense, responsable, ordre*) par exemple.

3.1.3 Bilan

À l'issue de l'exposé de cette première version d'implémentation de notre méthodologie d'acquisition de lexiques basés sur la SD, nous pouvons dresser un bilan des aspects positifs et négatifs de celle-ci aux divers paliers, et pointer les aspects à perfectionner ou à développer.

Concernant la phase de caractérisation des thèmes à l'aide de listes de mots, deux problèmes subsistent. D'une part, lors de la production de l'arbre de clas-

sification par CHAVL, un certain nombre de paramètres doivent être réglés pour aboutir à un arbre dont l'équilibre général laisse penser que la classification est satisfaisante. Peu d'indices (à part des expérimentations successives...) existent pour déterminer de manière très précise ces valeurs ; or nous avons constaté que de très légères variations de celles-ci entraînaient la production d'arbres de qualités très différentes. Cette dépendance est due au fait que `chavleps` manipule des données très peu denses (chacun des 165 N étudiés n'apparaissant que dans très peu des 9500 paragraphes). De plus, les cases non vides du tableau de contingence croisant les lemmes des N et les numéros de paragraphes, sont très faiblement chargées, puisque leur contenu représente le nombre d'occurrences du nom considéré dans le paragraphe désigné. Il convient donc de trouver une méthode de densification de ce tableau. D'autre part, les classes effectivement porteuses de thèmes doivent être désignées manuellement dans l'arbre, le critère de l'AVL limitant éventuellement les propositions mais ne résolvant pas ce problème. Enfin, il faut étudier la qualité de détection (précision) que les listes caractérisant les thèmes permettent d'atteindre et accroître au maximum leur couverture du corpus. La section 3.2.1 présente les moyens que nous avons mis en œuvre pour répondre à ces objectifs.

Pour ce qui est de la phase de création des taxèmes et de mise au jour de sèmes, le travail à réaliser est encore important. Si nous avons momentanément mis de côté les difficultés de production de classes sémantiquement homogènes en nous basant sur la faisabilité (plus ou moins attestée quant à l'effort manuel nécessité pour son exploitation effective) démontrée de cette tâche dans le cadre de corpus spécialisés, nous allons devoir, une fois que l'extraction des sous-corpus thématiques sera avérée, travailler sur la qualité de la méthode de classification choisie et des contextes exploités pour ce faire. Enfin, concernant l'analyse sémique, si le travail manuel d'extraction de séquences au sein des contextes syntagmatiques des membres d'une même classe a montré que l'on pouvait faire émerger des aspects de sens pertinents, notre objectif va, entre autres, être d'automatiser autant que possible la sélection de ces séquences. Nous abordons en section 3.2.2 différentes pistes que nous voulons explorer pour ces deux tâches.

3.2 Perfectionnement des étapes

Nous abordons ici successivement les travaux que nous avons menés pour améliorer la qualité et l'automatisation de la détection des thèmes, puis nous discutons nos recherches à venir sur la détermination et la structuration des taxèmes au sein de corpus spécialisés, ainsi que quelques perspectives à un peu plus long terme.

3.2.1 Caractérisation des thèmes

Les conditions d'expérimentation des travaux décrits dans cette section sont les suivantes : pour augmenter la qualité de celui-ci, nous avons, dans un premier temps, repris l'étiquetage du corpus du *Monde diplomatique*, à l'aide des outils Multext MtSeg de l'Université de Provence [IV94] et Mmorph de l'Issco [Arm96, PR94], la désambiguïsation étant toujours effectuée par le logiciel Tatoo. De plus,

contrairement à la première expérience, nous avons pris en compte l'intégralité de ce corpus (archives de 1984 à 1998), soit 11,4 millions de mots répartis en 98 000 paragraphes. Nous en avons extrait un échantillon aléatoire de 8000 paragraphes (700 000 mots environ) et avons retenu les 383 N apparaissant au moins 60 fois dans cet extrait.

Comme nous l'avons mentionné précédemment, lors de la classification effectuée par `chavleps`, la faible densité des données manipulées rend la qualité de l'arbre obtenu extrêmement fluctuante et dépendante de faibles écarts dans le choix des valeurs de certains paramètres. Les caractéristiques du tableau de contingence manipulé – 98% de cases vides pour cette matrice comportant en ligne les 383 lemmes et en colonnes les 8000 numéros de paragraphes, et cases non vides peu chargées – font que certaines mesures statistiques élaborées employées par CHAVL ont des valeurs très faibles qui disparaissent à cause d'approximations dues à des calculs sur des valeurs à virgule flottante. Il est donc nécessaire de densifier cette matrice de répartition des mots.

De plus, nous avons besoin d'un mode de lecture de l'arbre qui permette non seulement de pouvoir déterminer automatiquement les classes les plus pertinentes à différents niveaux (la partition à un seul niveau n'étant pas satisfaisante), mais également d'opérer, quand nécessaire, des réorganisations mineures de cet arbre pour éviter l'insertion de quelques « intrus » dans des classes thématiquement homogènes, voire réinsérer de tels éléments dans une autre classe où ils seraient plus pertinents. Ainsi, dans l'arbre de classification de la figure 3.3 en page 47, on aimerait pouvoir déplacer *ville* vers la classe {*logement*, *cité*}.

Nous avons apporté une solution à ce double problème à l'aide d'un « outil » commun : une seconde classification effectuée sur les paragraphes du corpus d'apprentissage en fonction des mots qu'ils partagent. Cette classification de paragraphes sert d'une part à effectuer la réduction souhaitée des données pour la classification des mots en substituant au tableau de contingence des noms et paragraphes un tableau de contingence des noms et classes de paragraphes. La partition des paragraphes est également mise à profit pour développer une méthode de recherche de classes optimales, en confrontant la qualification thématique opérée par les classes de mots à la partition des paragraphes, utilisée comme référent, grâce à une mesure évaluant la corrélation existant entre ces deux réponses apportées au même problème de la reconnaissance de thèmes. Nous présentons ici les principes de cette solution décrite en détail dans [RS02].

Pour alléger notre propos, nous parlons dans la suite de p-classification (p-classe, p-partition) lorsque nous mentionnons la classification effectuée sur les paragraphes et de m-classification (m-classe, m-partition) lorsque nous faisons référence à la classification des noms en fonction de leur distribution dans les paragraphes.

Classification des paragraphes

Mesure de similarité des paragraphes Pour cette p-classification, nous avons choisi une mesure de similarité qui reflète le concept de cohésion lexicale évoqué en section 3.1.1, et qui consiste à déterminer le nombre de mots partagés par deux paragraphes pour évaluer leur proximité thématique. Nous affinons toutefois ce principe

en donnant un poids supplémentaire au partage de mots rares, dont l'apparition vraisemblable dans peu de thèmes distincts fait de leur coprésence dans deux paragraphes un indicateur fort pour la classification thématique. L'importance de chaque mot est donc inversement proportionnelle à son nombre d'occurrences dans le corpus d'apprentissage, et la mesure est normalisée en fonction de la taille des paragraphes comparés. Par conséquent, la similarité entre deux paragraphes A et B est définie par :

$$\frac{1}{\min(p,q)} \sum_i \frac{\min(a_i,b_i)}{n_i}$$

où $A = (a_i)$ et $B = (b_i)$ sont les vecteurs rassemblant le nombre d'occurrences de chaque mot considéré pour ce calcul dans chacun des paragraphes, n_i est le nombre total d'occurrences du mot i dans le corpus d'apprentissage, et p et q les nombres de mots dans les deux paragraphes.

Partition des paragraphes Cette mesure pouvant prendre en compte les mots rares nous permet d'effectuer un calcul de similarité pour tous les N apparaissant au moins 2 fois, soit 3000 N, et de compenser par un grand volume de données la relative simplicité de la mesure. La demi-matrice 8000x8000 [4] contenant les valeurs de similarité entre paires de paragraphes est utilisée par `chavleps` pour construire un arbre de p-classification. Cet arbre, bien équilibré, nous permet d'extraire, par coupure à un niveau de l'arbre où la taille moyenne des classes atteint une valeur que nous avons choisie à 12 [5], un ensemble de 544 classes.

Cette partition des paragraphes ne peut cependant être considérée comme un résultat en soi pour notre problème de détection de thèmes dans le corpus, mais doit uniquement être vue comme une étape permettant d'améliorer la m-classification. Plusieurs raisons expliquent ce fait : d'une part, si des p-classes tirées aléatoirement montrent un certain niveau de consistance thématique, leur qualité n'est pas suffisante pour qu'elles soient considérées comme un résultat final ; de plus, rien n'assure que tous les paragraphes traitant d'un thème donné soient réunis en une unique p-classe ; par ailleurs, le calcul de la mesure de similarité entre tous les couples de paragraphes d'un corpus conséquent serait, pour des raisons calculatoires, impossible ; enfin, la p-partition ne donne pas d'information sur le thème reconnu alors que notre technique à base de listes de mots permet un résultat interprétable.

Exploitation de la p-classification

Densification de la matrice de répartition Les p-classes servent tout d'abord à répondre au besoin de densification de la matrice de répartition utilisée lors de la m-classification. Nous passons en effet d'un tableau de contingence croisant 383 N et 8000 numéros de paragraphes à un tableau, beaucoup moins creux, croisant 383 N et 544 classes de paragraphes. Son traitement par `chavleps` conduit à un

4. La mesure est symétrique.

5. Cette valeur empirique est un bon compromis entre généralisation et consistance thématique des p-classes.

arbre mieux équilibré, dans lequel la plupart des fusions suggérées, certes non encore parfaites, sont intuitivement satisfaisantes. De plus, les petites variations des valeurs de paramètres ont maintenant un effet négligeable sur l'arbre produit.

Nous utilisons également la p-classification pour définir une mesure de qualité des classes de mots et nous guider dans le choix des m-classes proposées par l'arbre de m-classification ou légèrement modifiées.

Définition d'une mesure de qualité des classes Les classes de mots porteuses de thèmes que nous voulons obtenir ont pour objectif d'être utilisées d'une façon que nous avons déjà mentionnée, à savoir : si au moins deux mots d'une liste caractérisant un thème sont présents dans un paragraphe du corpus, alors ce paragraphe évoque ce thème. On dira par la suite que la m-classe « reconnaît » le paragraphe. Par conséquent, l'ensemble des m-classes que nous voulons extraire de l'arbre de m-classification effectue une classification thématique des paragraphes étudiés, ce qui est aussi le rôle de la p-partition que nous venons de définir. Ces deux classifications devraient donc coïncider autant que possible[6]. Si nous considérons le cas idéal où tous les paragraphes d'une p-classe concernent un même thème et où chaque m-classe reconnaît l'ensemble des paragraphes évoquant le thème qu'elle caractérise (et seulement eux), alors une m-classe reconnaît soit tous les paragraphes d'une p-classe, soit aucun. La mesure de qualité définie donne la préférence aux m-classes les plus proches de cette configuration idéale.

Soit $\mathcal{M}$ une m-classe et $\mathcal{P}_1, \ldots, \mathcal{P}_n$ toutes les p-classes définies par la p-partition ($n = 544$). Pour chaque paragraphe P, rec $(\mathcal{M},P)$ exprime le fait que $\mathcal{M}$ reconnaît P. On associe à $\mathcal{M}$ un vecteur $M \in \mathbb{R}^n$ défini par :

$$M = (m_1, \ldots, m_n), \text{ avec } \forall i \in [1,n]\,,\ m_i = \frac{\text{Card}\,\{P \in \mathcal{P}_i \mid \text{rec}\,(\mathcal{M},P)\}}{\text{Card}\,(\mathcal{P}_i)}$$

Chaque élément m_i de M représente donc la proportion de paragraphes de $\mathcal{P}_i$ reconnus par $\mathcal{M}$. Comme la numérotation des p-classes est totalement arbitraire, on peut définir $M' \in \mathbb{R}^n$, $M' = (m'_1, \ldots, m'_n)$ un vecteur contenant les mêmes valeurs que M mais classées par ordre décroissant. Nous dérivons notre mesure de qualité du profil global de ce vecteur.

La figure 3.2 donne (de manière simplifiée) diverses possibilités pour ce profil. Le premier cas correspond à une m-classe assez proche du cas idéal recherché (dans lequel les valeurs sont 0 ou 1) : il existe une séparation claire entre les 4 premières p-classes, dont beaucoup de paragraphes sont reconnus, et les autres. Dans le deuxième, il y a certes des différences de taux de reconnaissance des paragraphes des diverses p-classes, mais ces p-classes sont difficiles à scinder en deux catégories. Le troisième cas est le pire, dans lequel la m-classe n'opère aucune distinction entre les p-classes.

Pour détecter et distinguer ces divers cas, la mesure de qualité que nous avons définie est basée sur la « dérivée » $M'' \in \mathbb{R}^{n-1}$ de M' (*cf.* ligne du bas de la

6. La correspondance est essentiellement limitée par le fait qu'un paragraphe peut être reconnu par plusieurs m-classes mais n'appartient qu'à une seule p-classe.

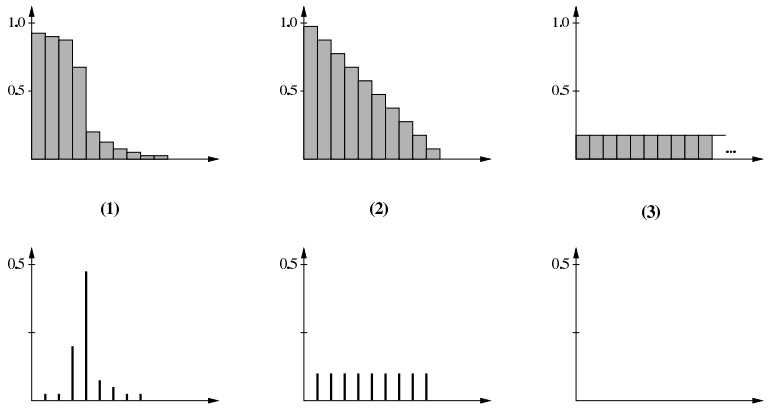

FIG. 3.2 – *En haut, trois graphiques montrant, pour une collection de p-classes (axe des x) quelle proportion de leurs paragraphes est reconnue par une m-classe donnée (axe des y). En bas, différences entre les valeurs consécutives de proportion.*

figure 3.2), soit :

$$M'' = (m''_1, \ldots, m''_{n-1}), \text{ avec } \forall i \in [1, n-1]\,,\ m''_i = m'_{i+1} - m'_i$$

L'explication de son expression exacte est donnée dans [RS02] et nous nous limitons ici à sa formulation.

$$q\,(\mathcal{M}) = (1 + \sigma_{M''})\,(1 + (m'_1 - m'_n)) - 1$$

où $\sigma_{M''}$ est l'écart-type des valeurs de ce vecteur.

Lecture de l'arbre de classification Cette fonction q est exploitée par un algorithme pour permettre une lecture « intelligente » et guidée de l'arbre de m-classification initial. D'une part q sert à pointer les classes pertinentes dans l'arbre quel que soit leur niveau. D'autre part, elle permet d'ignorer certaines fusions effectuées par `chavleps`, voire de les modifier. Nous présentons le déroulement de cet algorithme en nous appuyant sur un exemple (*cf.* figure 3.3). L'algorithme part des feuilles de l'arbre de m-classification et remonte vers la racine en vérifiant si, à chaque nœud, les fusions proposées accroissent la valeur de q.

- Si c'est le cas, l'algorithme effectue la fusion et continue l'exploration ascendante de l'arbre avec cette nouvelle classe.
- Sinon, l'algorithme continue à remonter vers la racine mais sans effectuer la fusion. On se retrouve donc avec un ensemble de classes au lieu d'une m-classe. Par exemple, après (b), il y a un ensemble de 2 classes {{*cinéma*, *film*, *scène*}, {*ville*}}.

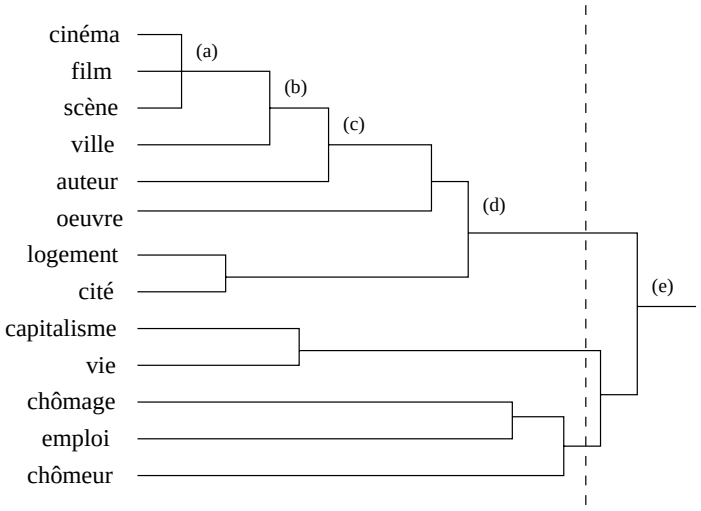

FIG. 3.3 – *Exemple réduit d'arbre de classification des noms produit par* `chavleps`. *La ligne pointillée indique un niveau de lecture traditionnel de tels arbres.*

Aux nœuds supérieurs, toutes les possibilités de fusions entre membres des ensembles de classes sont testées pour détecter celle qui est la plus intéressante en termes d'évolution de q :

- En (c): {{*cinéma*, *film*, *scène*}, {*ville*}} « + » {*auteur*}
  → {{*cinéma*, *film*, *scène*, *auteur*}, {*ville*}}
- En (d): {{*cinéma*, *film*, *scène*, *auteur*, *œuvre*}, {*ville*}} + {*logement*, *cité*}
  → {{*cinéma*, *film*, *scène*, *auteur*, *œuvre*}, {*ville*, *logement*, *cité*}}

Finalement, nous obtenons en (e), racine de l'arbre, la partition : {{*cinéma*, *film*, *scène*, *auteur*, *œuvre*}, {*ville*, *logement*, *cité*}, {*capitalisme*,*vie*}, {*chômage*, *emploi*, *chômeur*}}.

Les classes finales sont donc différentes de celles produites par `chavleps` et sont obtenues à l'issue du parcours sans qu'il soit nécessaire de les chercher à divers niveaux de l'arbre. Un certain nombre d'heuristiques permettent d'avoir des temps de calcul intéressant et de filtrer quelques classes peu utiles (telles que {*capitalisme*,*vie*}). Ces détails peuvent être trouvés dans [Ros01].

Résultats intermédiaires

En appliquant cette méthode de classification à notre corpus d'apprentissage, nous obtenons, à l'issue de la lecture « intelligente » de l'arbre de m-classification, 36 classes dont 25 présentent une cohérence nette (ce résultat est à rapprocher des proportions bien moindres de la première version (21/45 ou 27/80)), par exemple {*bureau, centre, enseignement, institution, recherche, université, école*} ou {*chaîne, image, information, moyen, média, programme, réseau, télévision, événement*}. Les autres ne sont pas réellement dénuées de consistance sémantique mais reflètent

plutôt des concepts un peu vagues que ces classes capturent de manière trop lâche pour que l'on puisse les considérer comme pertinentes pour une tâche de détection de thèmes (par exemple {*communauté, dimension, esprit, mesure, mouvement*}). En confrontant les thèmes que ces classes révèlent à la connaissance d'un lecteur régulier du mensuel, nous pouvons également noter que les classes détectent les thèmes principaux du corpus.

Une validation plus quantifiable est cependant réalisable. Elle consiste à estimer la qualité des classes par rapport à la tâche qu'elles ont à effectuer. Ces classes ont pour but de permettre de scinder au mieux le corpus initial en sous-corpus thématiquement homogènes, afin d'avoir un matériau linguistique suffisant et spécialisé pour appliquer la suite de notre méthodologie et produire des lexiques sémantiques basés sur la SD. On peut donc s'intéresser à la précision de la répartition des paragraphes dans les thèmes qu'elles permettent d'atteindre d'une part, et à leur couverture du corpus d'autre part (c'est-à-dire la proportion des paragraphes qu'elles permettent effectivement de répartir).

Si l'on utilise les 25 classes pour détecter, à l'aide de la coprésence d'au moins deux de leurs membres, les thèmes abordés dans l'ensemble du corpus, la précision, évaluée sur 1000 paragraphes tirés aléatoirement, est de l'ordre de 55%, et la couverture n'est que d'un tiers des paragraphes ; ceci s'explique certes par le fait que la classification n'a été effectuée que sur 383 N (qui fait que les classes obtenues sont petites et que certains thèmes mineurs ne sont pas atteignables), mais ce résultat doit être amélioré. Une solution envisageable serait, par exemple, de combiner les classes obtenues et des indices linguistiques, permettant d'étendre un thème reconnu dans deux paragraphes distants d'un même article aux paragraphes intermédiaires si aucune scission évidente du discours n'y a été détectée. Ceci permettrait d'accroître la couverture, mais pas la précision. Nous avons plutôt choisi d'utiliser une combinaison d'exécutions de la méthode de détection de classes de lexies symptomatiques de thèmes, pour mettre au jour des noyaux communs permettant de bâtir des classes plus étendues et plus fiables. Cette méthode repose sur les principes que nous venons de détailler ici et nous la décrivons donc brièvement pour terminer cette section.

Combinaison d'exécutions : résultats et évaluation

Description de la méthode À partir du corpus initial du *Monde diplomatique*, nous exécutons n fois[7] la procédure de recherche de mots-clés précédente : nous tirons aléatoirement 10 000 paragraphes ; nous en extrayons les 1000 N les plus fréquents[8] ; pour la classification des paragraphes, tous les noms apparaissant au moins 2 fois sont pris en compte (6000 à 6500) ; nous exécutons la p-classification, la m-classification et la lecture de l'arbre à l'aide de la fonction q.

L'« intersection » des n ensembles de classes obtenus se fait en représentant les associations de N par un graphe valué, dont les nœuds sont les N présents dans

7. Au-delà de 30 itérations, chacune d'une durée de 15 à 20 minutes sur une station Sun Blade, il y a stabilisation du résultat.

8. Si ce nombre élevé de mots a tendance à rendre les premières classes produites moins homogènes, la suite du processus appliqué permet d'aboutir à une qualité très satisfaisante.

au moins une classe, et dont le poids d'un arc entre deux N correspond au nombre de fois où ces deux N ont été réunis dans une classe. On itère alors le traitement suivant : repérage de l'arc de poids le plus fort ; définition d'un palier juste en-dessous du poids de cet arc, sous lequel on considère les autres arcs comme inexistants ; recherche de la composante connexe du graphe comportant l'arc de poids maximal en ajustant le palier pour que cette composante connexe contienne entre 4 et 10 mots ; récupération du noyau de classe ainsi sélectionné et extraction de ces nœuds du graphe. À l'issue de ces itérations, on obtient une quarantaine de noyaux de 5 ou 6 mots en moyenne.

Ensuite, ces noyaux sont étendus en parallèle, en travaillant sur la totalité du corpus, de la façon suivante : on considère le sous-ensemble des paragraphes reconnus par chaque noyau ; on ajoute au noyau les mots dont la fréquence sur ce sous-ensemble de paragraphes est particulièrement élevée par rapport à leur fréquence moyenne sur le corpus ; à chaque ajout de mot, l'ensemble des paragraphes reconnus est recalculé ; la procédure d'agrégation s'arrête lorsque l'ajout d'un mot fait chuter le nombre moyen de mots symptomatiques de thèmes par paragraphe reconnu. On obtient une quarantaine de classes de 25-30 mots, telles que {*alcool, argent, baron, cartel, circuit, cocaïne, contrebande, corruption, criminalité, destination, drogue, délinquance, enlèvement, gang, héroïne, mafia, meurtre, opium, trafic, viol, vol*} et {*champion, chaussure, club, compétition, exploit, football, foule, gloire, joueur, match, performance, pilote, plume, report, sport, stade, tribune, vedette*}, qui sont intuitivement toutes très satisfaisantes en termes de consistance thématique.

Évaluation - validation des classes En utilisant les critères précédents de reconnaissance d'un thème dans un paragraphe à l'aide de deux mots de sa classe caractérisante, ces nouvelles m-classes permettent de couvrir deux tiers des paragraphes du corpus – soit 70% du corpus en nombre de mots – , avec une précision de l'ordre de 55%. Nos classes plus vastes nous permettent cependant d'accroître la précision en utilisant la coprésence de 3 mots-clés comme indice thématique. Dans ce cas, la précision atteint 85%, la couverture redescendant quant à elle au tiers des paragraphes, soit environ 40% du corpus en nombre de mots. Nous pouvons toutefois perfectionner ce critère en prenant en considération la structure des articles de notre corpus et en utilisant, dans les articles où le critère « 3 mots-clés » a permis de sélectionner au moins deux paragraphes, la coprésence de 2 mots de la liste caractéristique du même thème pour chercher à reconnaître leurs autres paragraphes. Pour contrôler la précision, nous appliquons parallèlement un « durcissement » du critère de sélection pour les articles où un seul paragraphe a été reconnu comme évoquant un thème par le critère « 3 mots-clés » : dans ce cas, nous ne retenons effectivement le paragraphe que s'il contient 4 mots caractérisant le thème. Cette combinaison de critères permet d'atteindre une couverture de l'ordre de 58% (65% en nombre de mots) avec une précision de 85%. Pour information, une rapide étude des paragraphes non « classés » thématiquement nous a conduite à constater qu'environ un quart de ceux-ci est effectivement très difficilement classable, puisque que ces paragraphes peuvent être deux lignes de transition, une très rapide notice biographique d'une personne interviewée...

3.2.2 Constitution et structuration de taxèmes

Contrairement à la phase précédente de la méthodologie d'acquisition de relations intracatégorielles basées sur la SD, nous n'avons pas encore eu le temps de porter nos efforts sur le perfectionnement de la production de taxèmes et la structuration de ceux-ci à l'aide de sèmes spécifiques. Dans cette section, nous allons donc présenter des recherches que nous voulons mener dans cette optique, certaines dans un avenir très proche et d'autres à un peu plus long terme.

Nous abordons, dans un premier temps nos travaux à venir concernant la constitution automatique de classes sémantiques à partir de corpus thématiquement homogènes, obtenus grâce aux listes caractéristiques de thèmes précédemment décrites. Nous expliquons ensuite les pistes que nous allons explorer pour mettre au jour une organisation sémique de ces taxèmes à partir des premiers résultats décrits en section 3.1.2. Pour cette double présentation, nous nous appuyons sur [BHNZ97] et [FHL97], travaux qui abordent des idées qui sont les plus proches des nôtres, et dont nous voulons élargir et systématiser certaines propositions, émises à partir d'une analyse manuelle de regroupements automatiques basés sur des partages de contextes syntaxiques. Nous souhaitons également vérifier si les méthodes employées pour faire émerger des liens sémiques entre mots permettent de constituer également des relations sémantiques plus communément utilisées, telles que la synonymie, l'hyperonymie... et présentons quelques réflexions sur le sujet. Nous terminons par l'évocation d'une perspective à plus long terme : l'étude de la variation de sens induite par le remplacement d'un mot par une lexie à laquelle il est lié par des sèmes spécifiques, qui a un intérêt évident, en particulier dans des cadres applicatifs (extension de requêtes en RI, système d'aide à la génération de textes...).

Constitution de taxèmes

Comme nous l'avons déjà mentionné, tant en section 2.1.2 qu'en 3.1.2, de nombreux travaux ont été dédiés à la constitution de classes sémantiques à partir de corpus, en prenant pour voisinages des mots sur lesquels baser les regroupements, soit des contextes syntaxiques, soit des éléments apparaissant dans des fenêtres graphiques autour des mots étudiés (*cf.* par exemple [Gre94b, Aga95, Ass98]).

Lorsque ces travaux sont appliqués à des corpus spécialisés, les classes auxquelles ils permettent d'aboutir, même si elles requièrent parfois une intervention humaine, peuvent être considérées comme des groupes sémantiques homogènes. C'est par exemple le cas dans [BHNZ97], où les regroupements fournis peuvent être mis en relation avec des classes de concepts du domaine du corpus étudié.

Puisque nous allons à plusieurs reprises faire mention des travaux présentés dans [BHNZ97] et [FHL97] dans cette partie, nous nous attardons quelque peu pour les décrire ici afin de faciliter la compréhension du lecteur. [BHNZ97] expose et étudie les classes obtenues à partir d'un corpus médical (maladies coronariennes) à l'aide du logiciel `Zellig`. Celui-ci regroupe les mots sur la base de contextes syntaxiques normalisés. Plus précisément, il s'appuie sur les arbres d'analyse produits par des extracteurs de groupes nominaux et en dérive des dépendances binaires entre deux mots pleins (N ou A), de type tête-modifieur ou tête-argument. Il forme

ensuite un graphe dans lequel les nœuds correspondent aux lemmes et les arêtes sont étiquetées par les contextes partagés. Seuls les arcs indiquant le partage d'un nombre conséquent de contextes sont conservés. [BHNZ97] étudie les cliques et composantes connexes du graphe sur le seul corpus médical, alors que [FHL97] compare ces premiers résultats à ceux obtenus à partir d'un corpus non spécialisé, celui des discours radio-télévisés du premier septennat de François Mitterrand. Les composantes connexes du graphe obtenu sur le corpus spécialisé forment une catégorisation conceptuelle, alors que les résultats sur le second corpus sont beaucoup moins intéressants sur ce plan.

Pour notre part, nous partons d'un corpus général, mais l'utilisation des classes de lexies caractérisant les thèmes qu'il aborde permet de le scinder en sous-corpus thématiquement homogènes. Jusqu'à présent, pour mettre au jour des classes sémantiques au sein de chacun de ces sous-corpus spécialisés, nous avons utilisé une méthode très simple consistant à regrouper, par classification hiérarchique, les noms dont les vecteurs de contexte, formés de N, V et A apparaissant dans une fenêtre de plus ou moins 5 mots, sont les plus similaires (*cf.* section 3.1.2). Nous avons obtenu des classes, certaines très bruitées, d'autres plus homogènes, et nous sommes limitée à explorer quelques-unes de ces dernières, choisies manuellement.

Nos perspectives, dans ce domaine, vont consister, tout en gardant un contexte varié, c'est-à-dire non limité à des groupes nominaux par exemple, à passer d'un voisinage « sac de mots » où les éléments sont tous considérés de la même façon, à un contexte plus affiné dans lequel nous allons distinguer les positions (droite ou gauche) d'apparition des composants par rapport à la lexie cible et leur distance par rapport à celle-ci. Une première expérimentation en ce sens a été réalisée [Tar00], mais des critères formels de pondération des différents éléments participant à de tels contextes lors des calculs de similarité restent encore à établir. Ceci permet, outre la comparaison de vocabulaire, de définir une mesure de similarité prenant en compte, si nécessaire, la ressemblance des successions de catégories morpho-syntaxiques. [FHL97] mentionne d'ailleurs l'importance de certaines structures telles que *N de N* pour reconnaître des synonymes dans le cadre du traitement des seuls groupes nominaux élémentaires. Notre proposition étend donc cette idée à des contextes plus variés, puisque nous n'avons aucun *a priori* sur les éléments de voisinage pertinents pour mener ensuite à bien une analyse sémique. Nous allons également, comme dans le cas de la caractérisation des thèmes, essayer de spécialiser notre algorithme de classification pour tenter d'obtenir des taxèmes suffisamment fiables de la façon la plus automatique possible.

Analyse sémique

Dans le premier travail effectué sur la structuration des taxèmes par des sèmes spécifiques exposé en section 3.1.2, nous avons montré que l'exploration manuelle des ensembles formés par les différences de voisinages des mots regroupés dans une même classe permettait de mettre au jour des séquences de mots caractérisant un trait sémantique d'une lexie par rapport à une autre. La même recherche menée sur les contextes d'un même mot dans différents thèmes conduisait, quant à elle, à

pointer des facettes de sa signification[9].

Sur le corpus médical, [BHNZ97] montre également qu'une exploration manuelle et un typage des éléments contextuels des lexies fortement liées dans le graphe produit par `Zellig` permettent de révéler des informations sémantiques intéressantes. Ainsi, dans les voisinages que *sténose* entretient avec de nombreux mots, cinq groupes de propriétés de ce terme peuvent être caractérisées : son aspect, sa gravité, sa localisation, l'acte thérapeutique et sa connotation processus. Des lexies partageant avec lui tous ces types de contexte ont tendance à en être des synonymes, alors que celles qui partagent uniquement certains d'entre eux, et dans certaines proportions, forment des groupes sémantiques homogènes liés à *sténose* par diverses autres relations de sens. Dans [FHL97], l'étude manuelle de quelques cliques et composantes connexes obtenues à partir du corpus non spécialisé `Mitterrand` fait apparaître des connexions de quelques lexies avec plusieurs ensembles homogènes de mots qui mettent chacun en évidence une de leurs facettes de sens.

Notre objectif est de systématiser l'exploitation des contextes partagés ou non par des mots au sein d'un taxème et par un même mot à travers plusieurs thèmes pour faire émerger les séquences indicatrices de sèmes. Nous souhaitons déterminer des moyens d'automatiser cette tâche, et exploiter pour ce faire toute la connaissance à notre disposition, à savoir, par exemple, les contextes enrichis d'informations de distance et de position, le nombre d'occurrences des contextes partagés entre mots, le contexte spécifique d'un mot dans son taxème... Nous envisageons également d'explorer l'utilisation des séquences discriminantes entre les occurrences de mêmes mots dans des thèmes distincts, sur un nombre de mots suffisamment conséquent pour qu'une analyse statistique, si elle s'avère pertinente, soit fondée, afin de repérer des sous-séquences de mots qui ont une coprésence caractéristique à l'intérieur de ces différents thèmes.

Relations lexicales traditionnelles

Pour terminer cette section, nous nous intéressons à deux perspectives qui, si elles s'éloignent quelque peu de la « simple » production de lexiques sémantiques basés sur les principes de la SD, se basent directement sur l'analyse sémique et ont un intérêt fort dans des exploitations applicatives de ces lexiques.

Les « outils » que nous voulons mettre en place pour différencier les lexies par des sèmes spécifiques permettent-ils également de mettre au jour entre elles des relations lexicales plus traditionnelles telles que la synonymie, l'antonymie ou l'hyperonymie? Telle est la question sur laquelle nous voulons aussi nous pencher[10].

Nous avons déjà signalé que [BHNZ97] a constaté que les mots qui partagent tous les types de contextes d'un même mot ont tendance à en être des synonymes. Le partage limité à certaines propriétés uniquement permet également aux auteurs

9. [Fol02] partage cette idée de mise en évidence de particularités de sens de mots en contrastant en particulier, non pas à travers différents thèmes, mais à travers leur utilisation par six acteurs (direction et organisations syndicales) d'une même entreprise, leurs contextes extraits par `Zellig`.

10. Même s'ils peuvent peut-être nous servir de sources d'inspiration, nous sommes donc très éloignée des objectifs de travaux de mise au jour de structures syntagmatiques porteuses de telles relations, tels que [Mor99, Hea98] par exemple.

de pointer du doigt quelques hyperonymes. Nous étant attelée à faire émerger le plus automatiquement possible la visualisation de ces types par des séquences présentes dans les contextes partagés (ou non) par des lexies, notre objectif est alors de tenter de relier ces dernières et les sèmes qu'elles représentent à ce genre de relations lexicales, souvent très utiles dans des applications (*cf.* section 2.1.2). Nous avons déjà remarqué que la présence de certains mots d'un taxème dans le contexte spécifique d'un autre élément de la classe pouvait laisser penser à une structuration hyperonymique (*cf.* section 3.1.2). Même si cette idée simple peut servir de base de réflexion pour déterminer des moyens fiables de passage des « collections » de sèmes à de telles relations, nous pensons peut-être devoir employer des méthodes d'apprentissage, éventuellement supervisé [11], pour aboutir à une telle caractérisation.

Variations de sens

Faire émerger des sèmes liant les lexies d'un même taxème consiste à pointer des différences très fines de sens entre des mots proches sur l'axe paradigmatique. Ces différences, si nous nous limitons pour notre part à leur « visualisation » à l'aide de séquences de mots de contexte caractéristiques, peuvent donner lieu à une interprétation et un nommage. De telles relations à granularité fine fournissent certains moyens pour étudier ce que le remplacement d'un terme par un terme sémantiquement proche, par exemple dans le cadre d'une extension de requête en RI, risque d'induire comme variation de sens. Bien évidemment, cette question est particulièrement vaste puisqu'elle peut mener, selon le type d'application, jusqu'à devoir traiter de compréhension et plus particulièrement de la façon dont un utilisateur peut percevoir la modification. Nous nous limitons, déjà consciente de l'ampleur des interrogations soulevées avec cette restriction, au cadre de la seule interprétation linguistique.

Lors du remplacement d'un terme simple par un autre terme simple, la problématique soulevée ici peut conduire à se baser sur l'analyse sémique pour tenter, par exemple dans un cadre de RI, de caractériser en quoi les documents ramenés à l'aide du terme remplaçant seront différents de ceux retournés à l'aide du terme de base.

Si on s'intéresse à la substitution d'un constituant d'un terme complexe par une lexie sémantiquement proche dont on connaît les sèmes la liant au mot initial, la tentative de formalisation de la modification de sens induite devra vraisemblablement prendre en compte des aspects dynamiques de l'interprétation en SD que nous n'avons pas abordés jusqu'ici, c'est-à-dire d'activation et d'inhibition de sèmes entre composants du terme complexe. Des travaux comme ceux présentés dans [Beu98] peuvent fournir quelques pistes intéressantes.

Au même titre que les travaux sur les variations morpho-syntaxiques de termes de Jacquemin [Jac96] ou Daille [Dai00] par exemple, qui proposent des règles indiquant les conditions pour qu'une forme nouvelle soit une variante effective et licite d'un terme donné, on peut alors se demander s'il est possible de déterminer des

11. L'apprentissage supervisé, ou apprentissage à partir d'exemples, a pour but de construire des classifieurs (par exemple des règles de classification) à partir de données « étiquetées » par un expert, portant l'étiquette de leur classe.

contraintes, en termes de sèmes partagés ou non, pour que le remplacement d'une lexie par une autre poursuive l'évocation du « même » concept.

3.3 Conclusions

En guise de conclusion à ce chapitre, nous allons faire un bilan de nos contributions, d'une part par rapport à notre tâche première, c'est-à-dire l'apprentissage sur corpus de relations intracatégorielles basées sur la SD sans connaissance *a priori* – et donc le test du caractère implémentable de cette théorie, conduisant à envisager le développement de lexiques « à grande échelle » – , d'autre part sur un aspect plus lié aux techniques d'apprentissage.

Concernant le premier axe, même si nous sommes encore éloignée de la « mécanisation » complète, nous avons fait des propositions concrètes de mise en œuvre de lexiques sémantiques basés sur la SD à partir de corpus. Notre travail le plus accompli porte sur la mise au point d'une méthode fiable et automatique de détection et de caractérisation des thèmes présents dans un corpus non spécialisé. Les classes de mots auxquelles nous aboutissons sont vastes et homogènes, ce qui permet de repérer les thèmes effectivement abordés dans une partie conséquente du corpus et de disposer d'un matériau textuel suffisamment étendu pour la seconde phase. De plus, outre le rôle de palier intermédiaire dans le passage « corpus non spécialisé $\rightarrow$ lexiques basés sur la SD », ces caractérisations de thèmes peuvent trouver des intérêts applicatifs (filtrage...). Pour ce qui est de la seconde étape de constitution de taxèmes et de leur structuration par des sèmes spécifiques, nous nous attelons à une étude nouvelle et ambitieuse pour laquelle nous avons à présent essentiellement donné des pistes qui montrent une certaine faisabilité de la tâche. Le travail à réaliser étant encore conséquent, nous ne pouvons actuellement nous prononcer plus avant sur la part d'automatisation effective que l'on pourra atteindre, en particulier dans l'extraction de séquences de contexte caractérisant des sèmes.

La mise au point des listes de lexies caractéristiques des thèmes a été réalisée grâce à la définition d'une algorithmique originale de recherche de classification valide et signifiante dans le cas d'un « gros » tableau de contingence particulièrement creux. Grâce à un arbre de classification des paragraphes basé sur leur cohésion lexicale, nous avons densifié le tableau de contingence par regroupement des colonnes. De plus, l'ossature de cet arbre a fourni l'argument de la nouvelle algorithmique de remise en cause partielle des associations, en tenant compte d'indices spécifiques de discrimination de classes de paragraphes par une classe de noms.

À l'issue de l'exposé fait ici de nos travaux sur le sujet, on voit donc émerger des possibilités concrètes d'apprentissage de relations N-N basées sur les principes de la SD, et la théorie nous a servi de cadre fort pour mettre au point la méthode d'acquisition présentée. Avant de passer à la description de nos recherches concernant l'apprentissage de liens transcatégoriels N-V tels que définis dans la structure des qualia du Lexique génératif de Pustejovsky, nous pouvons noter que notre expérimentation de l'implémentation partielle[12] de la SD par l'apprentissage d'éléments

12. Comme nous l'avons déjà signalé, nous ne cherchons pas à implémenter l'intégralité des

pertinents en corpus fait émerger des échanges bilatéraux entre la théorie et l'empirique. D'une part, la théorie interpelle l'empirique puisque, sans que la SD ne fournisse obligatoirement des indices précis nécessaires, c'est par l'observation du corpus qu'il convient de faire effectivement émerger les éléments utiles pour implanter ses principes ; d'autre part, l'empirique peut lui aussi « poser des questions » à la théorie. Ainsi, lors de l'application sur le corpus de notre méthode de détection des thèmes, nous obtenons, par exemple, une classe de lexies caractérisant à la fois la peinture, le chant et la poésie, et des classes très focalisées, telles qu'une permettant de détecter tout ce qui a trait au FMI[13] et à la Banque mondiale, ou une autre, tout aussi homogène, symptomatique de l'économie monétaire. La question de la granularité des manifestations des isotopies génériques domaniales est donc posée par cette confrontation au corpus.

principes de la SD : nous nous cantonnons principalement au niveau microsémantique et ne mettons pas en œuvre l'aspect dynamique de la théorie, c'est-à-dire sa composante interprétative, qui décrit les activations et inhibitions de sèmes lors de la lecture d'un texte, d'une phrase ou d'un terme complexe.

13. Fonds monétaire international.

Chapitre 4

Apprentissage de relations nomino-verbales basées sur le Lexique génératif

Le Lexique génératif (LG) a déjà donné lieu à de nombreux travaux, dont la plupart porte cependant essentiellement sur la façon dont les représentations lexicales et les mécanismes génératifs de ce formalisme permettent d'exprimer certains phénomènes linguistiques (*cf.* [BB01] ou [BK01] par exemple). Très peu d'études s'intéressent réellement à la constitution « à grande échelle » de tels lexiques ; par exemple, le modèle *Simple* de Busa *et al.* [BCL01] présente uniquement le LG comme base de cadre unificateur de développement de lexiques dans différents langages, mais ne cherche pas à fournir des moyens d'automatiser la mise en place de telles ressources. Pour notre part, nous ne visons pas la constitution de lexiques génératifs complets mais avons pour objectif l'acquisition automatique sur corpus de couples N-V dont les constituants sont liés par un des rôles définis dans la structure des qualia, afin de pouvoir exploiter de telles ressources dans des cadres applicatifs. Ce chapitre présente une synthèse de nos travaux portant sur la mise au point d'une méthode d'apprentissage de telles paires N-V ; comme nous l'avons déjà signalé au chapitre 2, nous appelons par la suite ces paires couples ou paires qualia, par opposition aux couples N-V dont les constituants ne sont pas liés par un rôle qualia et qui sont qualifiés de non qualia.

De manière un peu simplifiée, on peut dire que deux options sont généralement prises en compte pour acquérir des couples N-V[1]. La première consiste à utiliser des approches statistiques, pour extraire des paires dont les constituants sont liés de façon statistiquement significative (*cf.* [Dai94] pour un tour d'horizon de ces méthodes) ; cependant, ces techniques ne sont pas suffisamment précises pour obtenir des relations ciblées (des paires N-V dont les membres sont liés par un des rôles qualia *versus* les autres paires dans notre cas). L'autre alternative est une approche

1. Plus généralement, pour acquérir des couples d'éléments quelconques.

linguistique d'extraction des couples N-V par repérage d'un ensemble de structures syntaxiques liées à l'expression de ces liens qualia ; c'est par exemple ce qui est proposé dans [PAB93], où les auteurs fondent l'acquisition de prédicats de la structure des qualia d'un mot sur quelques schémas syntaxiques caractéristiques des rôles qualia. L'avantage de tels patrons est la précision qu'ils permettent d'atteindre dans l'extraction. Toutefois, le problème majeur consiste à définir ces structures caractéristiques des relations qualia et ceci, pour tout nouveau corpus, question à laquelle le formalisme du LG ne répond pas. Nous proposons donc, pour notre part, d'aller un pas plus loin que cette approche linguistique, car nous n'avons pas d'*a priori* sur les structures susceptibles de porter des rôles qualia dans un corpus donné.

Pour acquérir en corpus des couples N-V liés par un lien qualia, nous avons choisi de développer une méthode d'apprentissage symbolique supervisé [2] de type programmation logique inductive (PLI) [MDR94], afin de produire des règles générales capables d'expliquer ce qui, en termes de contexte environnant, distingue les paires N-V qualia des autres paires. Ces règles sont ensuite appliquées sur le corpus étudié pour acquérir de nouvelles paires N-V qualia. Le fil conducteur de nos recherches, consistant, comme nous l'avons mentionné en introduction générale, à rendre l'apprentissage et les théories linguistiques compatibles, s'exprime ici par le fait que nous utilisons la théorie linguistique du LG pour préciser les paires N-V qui nous intéressent – ce ne sont pas « n'importe quelles » paires dont les constituants sont fortement liés que nous cherchons à acquérir, mais des paires dont le lien entre composants exprime un rôle qualia – , et pour évaluer et valider la qualité des couples que les règles apprises par PLI permettent effectivement d'acquérir. De plus, nous cherchons, en exploitant au mieux le caractère relationnel de la PLI, à apprendre des règles expressives et linguistiquement motivées, et non simplement des patrons efficaces pour l'extraction de paires qualia.

Dans cette partie introductive, nous expliquons et justifions le choix de ce cadre d'apprentissage, définissons certains critères auxquels nous souhaitons que notre méthode d'apprentissage réponde, puis présentons brièvement une vue globale des travaux que nous avons menés, afin d'expliciter notre démarche et les motivations de la version du système que nous présentons plus en détail dans les sections suivantes.

La PLI a pour but d'induire des théories – exprimées par des programmes logiques sous forme de clauses de Horn – à partir d'exemples et d'un ensemble de connaissances préalables (*background knowledge*). Plus précisément, un algorithme de PLI essaie de construire, à partir des connaissances préalables, des hypothèses génériques qui expliquent les exemples positifs (E^+) [3] tout en rejetant les exemples négatifs (E^-) (du moins le maximum d'exemples négatifs, un peu de bruit pouvant être toléré). Ces hypothèses sont le plus souvent générées en s'appuyant sur un langage (sorte de biais syntaxique), donné par l'utilisateur et qui assure ainsi la production d'hypothèses bien formatées par rapport au problème posé. Dans notre cas, les exemples positifs sont des paires N-V en contexte dont les constituants sont

2. Voir définition en note de bas de page numéro 11 page 53.

3. Sauf mention précise au sein d'une formule, nous notons indifféremment E^+ (respectivement E^-) l'ensemble des exemples positifs (respectivement négatifs), ou un ou plusieurs éléments de cet ensemble.

liés par l'un des rôles qualia (par exemple, *frein lâcher* dans *lâcher le frein de parking*). De même, nos exemples négatifs sont des couples N-V qui cooccurrent dans des phrases de notre corpus d'étude, mais sans que leurs éléments soient reliés par un tel rôle, par exemple *frein relâcher* dans *relâcher la commande de frein de parking*. Les hypothèses que nous cherchons à inférer à partir de ces E^+ et E^- sont des règles ou clauses, obtenues à l'aide d'un algorithme de PLI par généralisation contrôlée des E^+, et qui expliquent ce qui caractérise les couples qualia par rapport aux autres[4]. Ces règles, une fois appliquées sur le corpus, permettent d'extraire d'autres paires liées de la même manière. Ceci doit nous permettre de combiner la précision des schémas linguistiques pour les tâches d'extraction et la flexibilité d'une méthode automatique.

C'est le caractère explicatif de la PLI qui a motivé notre choix de ce cadre d'apprentissage : contrairement à des méthodes statistiques de type boîte noire donnant un résultat brut[5], elle ne se contente pas de construire un prédicteur (cette paire N-V est pertinente, celle-ci n'est pas qualia), mais elle fournit aussi une théorie fondée sur les données et infère des règles générales capables d'expliquer les exemples, c'est-à-dire d'apporter des éléments linguistiquement interprétables sur les relations qualia prédites. Le choix de la PLI est particulièrement justifié par le fait que la théorie du LG ne recense pas les structures syntaxiques exprimant ces relations et ne peut donc actuellement verbaliser des règles les décrivant, quel que soit le corpus. Nous cherchons sur ce point à contribuer à la théorie en offrant des règles explicatives et des structures porteuses de liens qualia dans les corpus étudiés. De plus, la PLI semble également une option appropriée car sa nature relationnelle peut fournir une expressivité puissante pour ces patrons linguistiques. Enfin, d'éventuelles erreurs d'étiquetage du corpus rendent essentielle la sélection d'une technique d'apprentissage tolérante aux fautes. La possibilité de traiter des données bruitées en PLI garantit la robustesse de la méthode d'apprentissage que nous développons.

Nous avons cherché à mettre au point une méthode d'apprentissage de couples N-V qualia qui respecte au mieux un certain nombre de critères. Elle doit évidemment être fiable, c'est-à-dire qu'elle a pour but de fournir des règles générales qui, une fois appliquées sur le corpus, permettent d'obtenir des couples N-V effectivement liés par un des rôles de la structure des qualia. De plus, comme nous l'avons déjà signalé, nous voulons que les règles produites soient linguistiquement motivées. N'ayant pas d'*a priori* sur les éléments de contexte adéquats pour caractériser les paires qualia, la méthode doit être capable de gérer de manière efficace une somme importante d'informations contextuelles pour en inférer des règles permettant d'extraire des paires pertinentes. Enfin, elle doit être réutilisable sur différents corpus à moindre coût, ce qui implique que sa mise en œuvre soit suffisamment légère et n'impose pas un prétraitement trop lourd du texte.

Nous présentons, dans ce chapitre, les travaux que nous avons réalisés pour essayer d'atteindre au mieux ces objectifs. Pour introduire la version la plus aboutie de notre méthode d'apprentissage de couples N-V qualia que nous décrivons dans les

4. Nous ne cherchons donc pas pour l'instant à distinguer les différents rôles qualia.
5. Éventuellement avec un coefficient d'association, mais sans explication.

sections suivantes, nous résumons brièvement ici la démarche globale que nous avons suivie et les différentes expériences d'apprentissage que nous avons réalisées. La première phase de mise au point d'une méthode d'apprentissage supervisé consiste à constituer les exemples positifs et négatifs du concept à apprendre. Dans un premier temps, nous avons formé ceux-ci à partir du contexte catégoriel des couples N-V retenus pour former ces exemples sur une partie (corpus d'apprentissage) de notre corpus d'étude, ainsi que sur des informations de distance entre le N et le V. La forme des exemples positifs dans ce cas était :

est_qualia(tag[6]_du_mot_avant_N, tag_du_mot_après_N, tag_du_mot_avant_V, tag_de_V, distance, position).[7]

où distance indique le nombre de verbes conjugués entre le N et le V dans la phrase et position spécifie si le V apparaît avant le N (codé pos[8]) dans la phrase ou le contraire (codé neg). Les exemples négatifs avaient la même forme. Le premier apprentissage a ainsi été réalisé, conduisant à la production de règles déjà pertinentes pour l'acquisition de couples N-V qualia [BFSJ00, SBF00, SBC+00]. Toutefois le seul étiquetage catégoriel du corpus ne permet pas de distinguer des cas où une même structure syntaxique indique parfois un couple qualia et parfois non. Par exemple, dans les structures du type « verb-inf det N1 prep N2 », le verbe à l'infinitif et le nom N2 ne sont parfois pas en relation (*vérifier l'absence de corrosion*) et le sont à d'autres occasions, notamment quand N1 indique un groupe (*vider les deux groupes de réservoirs*). Nous avons donc, dans un deuxième temps, ajouté un étiquetage sémantique aux mots de notre corpus (*cf.* section 4.1.2), et avons produit, pour l'apprentissage, des E^+ et des E^- de la forme :

est_qualia(tag_sémantique_précédant_N, tag_sémantique_suivant_N, tag_sémantique_précédant_V, tag_catégoriel_de_V, distance).,

où distance indique le nombre de verbes entre le N et le V, et, suivant son signe, l'ordre d'apparition du N et du V dans la phrase. C'est volontairement que nous n'exploitons pas toutes les informations disponibles dans le corpus, notamment les étiquettes sémantiques du N et du V, car nous voulons tester l'influence exacte de chaque type d'information. De plus, nos expériences successives d'apprentissage ont pour but d'introduire pas à pas de l'expressivité dans les règles, et ces deux premiers apprentissages n'exploitent pas, contrairement aux suivants, la puissance relationnelle de la PLI. L'apprentissage réalisé dans ce cas a conduit à de meilleurs résultats [BCFS01]. Nous avons dans un troisième temps pris en compte l'étiquetage sémantique du N et du V et de tous les mots qui cooccurrent avec eux dans la fenêtre de la phrase, et avons laissé l'algorithme d'apprentissage libre de trouver des généralisations « contraignant » les valeurs d'un nombre quelconque de ces éléments et non, comme dans les deux expériences précédentes, d'un nombre quelconque des éléments choisis comme champs fixes des clauses-exemples (6 dans la première et 5 dans la deuxième). Cette masse d'informations à traiter nous a conduite à travailler sur l'algorithme d'apprentissage pour régler, en particulier, des problèmes d'efficacité. Enfin, une quatrième expérience d'apprentissage a été réalisée en ne prenant plus en

6. Étiquette.
7. Le choix des éléments de contexte est expliqué dans [BCFS01, BFSJ00].
8. Pour positif, c'est-à-dire respectant l'ordre VN.

compte l'étiquetage sémantique des noms, étiquetage le plus lourd à réaliser ; nous cherchons, à ce niveau, à tester la portabilité de notre méthode. L'intégralité de ces expériences successives et de leurs résultats étant présentée dans [CSBF01], et les deux premières expériences ayant été réalisées dans des conditions très différentes des suivantes (format d'exemples différents, algorithmes de PLI différents...), nous nous focalisons ici sur la version la plus aboutie de notre méthode d'apprentissage et n'abordons donc que les deux dernières expériences. Nous nous référons toutefois aux premières pour justifier certains de nos choix, et pour montrer ce que les divers niveaux de prise en compte de l'étiquetage peuvent apporter.

Nous débutons la présentation de nos travaux par une description de notre corpus d'étude et de ses étiquetages morpho-syntaxique et sémantique. Les trois sections suivantes mettent successivement l'accent sur les critères d'efficacité, de fiabilité (à travers les validations tant théorique, empirique, que linguistique des clauses apprises) et de portabilité que nous cherchons à atteindre dans notre apprentissage de règles linguistiquement motivées d'extraction de couples N-V qualia. Ainsi, nous décrivons la méthode d'apprentissage développée en mettant l'accent sur la façon dont nous avons traité les problèmes de combinatoire inhérents à la prise en compte d'un volume conséquent d'informations contextuelles, et de production de règles significatives et expressives sur le plan linguistique ; nous détaillons ensuite les résultats obtenus en considérant l'intégralité de l'étiquetage d'une part, puis sans prendre en compte les étiquettes sémantiques des noms d'autre part. Nous terminons ce chapitre par un bilan de ce travail et des perspectives concernant, en particulier, l'application des couples N-V acquis dans le cadre d'applications de RI.

Plusieurs personnes ont collaboré très activement aux travaux décrits dans ce chapitre, dont Vincent Claveau qui a débuté une thèse sur ce sujet en octobre 2000 sous mon encadrement, Pierrette Bouillon de l'Issco (Genève) et Cécile Fabre de l'Erss (Toulouse) [9]. Laurence Jacqmin (FortisBank et département Infodoc de l'Université libre de Bruxelles) nous a également offert des cadres applicatifs pour de tout premiers tests de l'intérêt des couples N-V qualia pour l'extension de requêtes [10].

4.1 Le corpus et ses étiquetages

Le corpus que nous utilisons pour mettre au point notre méthode d'apprentissage de couples N-V qualia est un manuel de maintenance d'hélicoptères en français qui nous a été fourni par MATRA-CCR Aérospatiale. Il contient plus de 104 000 occurrences de mots, soit une taille d'environ 700 ko. Ce corpus a plusieurs caractéristiques intéressantes pour la tâche que nous visons : il est très homogène dans ses structures syntaxiques et son vocabulaire ; il compte un grand nombre de termes très concrets (*vis, porte...*) qui sont fréquemment utilisés au sein d'une phrase avec des verbes marquant leur rôle télique (*serrer les vis...*) ou leur rôle agentif (*effectuer*

9. D'autres personnes de l'Irisa nous ont fait partager leurs connaissances sur l'apprentissage automatique et la PLI, dont Jacques Nicolas et René Quiniou que nous tenons à remercier ici.

10. Ce quadruple partenariat Erss - Irisa - Issco - ULB a été initié dans le cadre d'une action de recherche partagée de l'Agence universitaire de la Francophonie (réseau Francil).

un réglage...).

4.1.1 Étiquetage catégoriel

Nous avons étiqueté catégoriellement ce corpus à l'aide d'outils développés dans le cadre du projet Multext [Arm96]. Le texte a donc été segmenté en unités lexicales ; les unités ont ensuite été analysées et lemmatisées, puis finalement désambiguïsées grâce à l'outil *Tatoo*, un étiqueteur à chaînes de Markov cachées [ABR95]. Ce traitement permet ainsi d'associer aux mots du corpus des informations – exploitées par la suite par notre système d'apprentissage – sur la catégorie morpho-syntaxique de chaque mot, et ce, avec une grande précision, puisque sur un échantillon-test de texte de 4 000 mots, seuls 2 % ont été détectés comme étant incorrectement étiquetés.

4.1.2 Étiquetage sémantique

Comme nous l'avons indiqué dans la partie introductive, afin de pouvoir distinguer les cas où une même structure syntaxique est parfois indicatrice d'un couple qualia et parfois non, nous avons réalisé un étiquetage sémantique du corpus MATRA-CCR (*cf.* [BCFS01] pour une présentation détaillée).

La méthode d'annotation sémantique que nous utilisons repose sur trois hypothèses majeures : (i) les informations catégorielles peuvent aider à distinguer les sens des mots polyfonctionnels [11] [WS96], (ii) l'étiquetage catégoriel (non ambigu) peut être réalisé par un étiqueteur probabiliste (voir 4.1.1), et, ce qui est plus surprenant, (iii) les ambiguïtés sémantiques restantes peuvent aussi être résolues par un étiqueteur probabiliste.

L'étiquetage sémantique du texte nécessite dans un premier temps de construire manuellement, pour chaque catégorie de mot, un lexique contenant pour chaque entrée les différentes étiquettes qu'elle peut porter au sein du corpus. Cela implique de choisir pour chaque catégorie un jeu d'étiquettes sémantiques adapté.

Pour classifier les noms du corpus de manière systématique, nous avons utilisé, comme point de départ, les classes les plus génériques de WordNet [Fel98]. Cependant, certaines de ces classes, inusitées dans notre corpus, ont été supprimées, alors que d'autres, très présentes, ont été raffinées en sous-classes plus précises (c'est le cas en particulier de la classe des objets concrets). Nous avons ainsi obtenu 33 classes, organisées en une hiérarchie représentée en figure 4.1 (les classes initiales de WordNet non usitées sont en italiques, et les étiquettes sémantiques choisies apparaissent entre parenthèses) [12]. Environ 8,7 % des entrées du lexique des noms constitué sont ambiguës. Ces ambiguïtés correspondent le plus souvent à des phénomènes de polysémie complémentaire (par exemple, *enfoncement* peut indiquer à la fois un processus et son résultat ; il est donc classifié en `pro` et `sta`).

11. C'est-à-dire qui ont plusieurs catégories, comme *règle* qui peut indiquer soit un nom, soit un verbe à l'indicatif.

12. Ce travail a été réalisé par Jean-Léon Bouraoui, étudiant en Maîtrise Sciences du langage de l'Université de Toulouse Le Mirail.

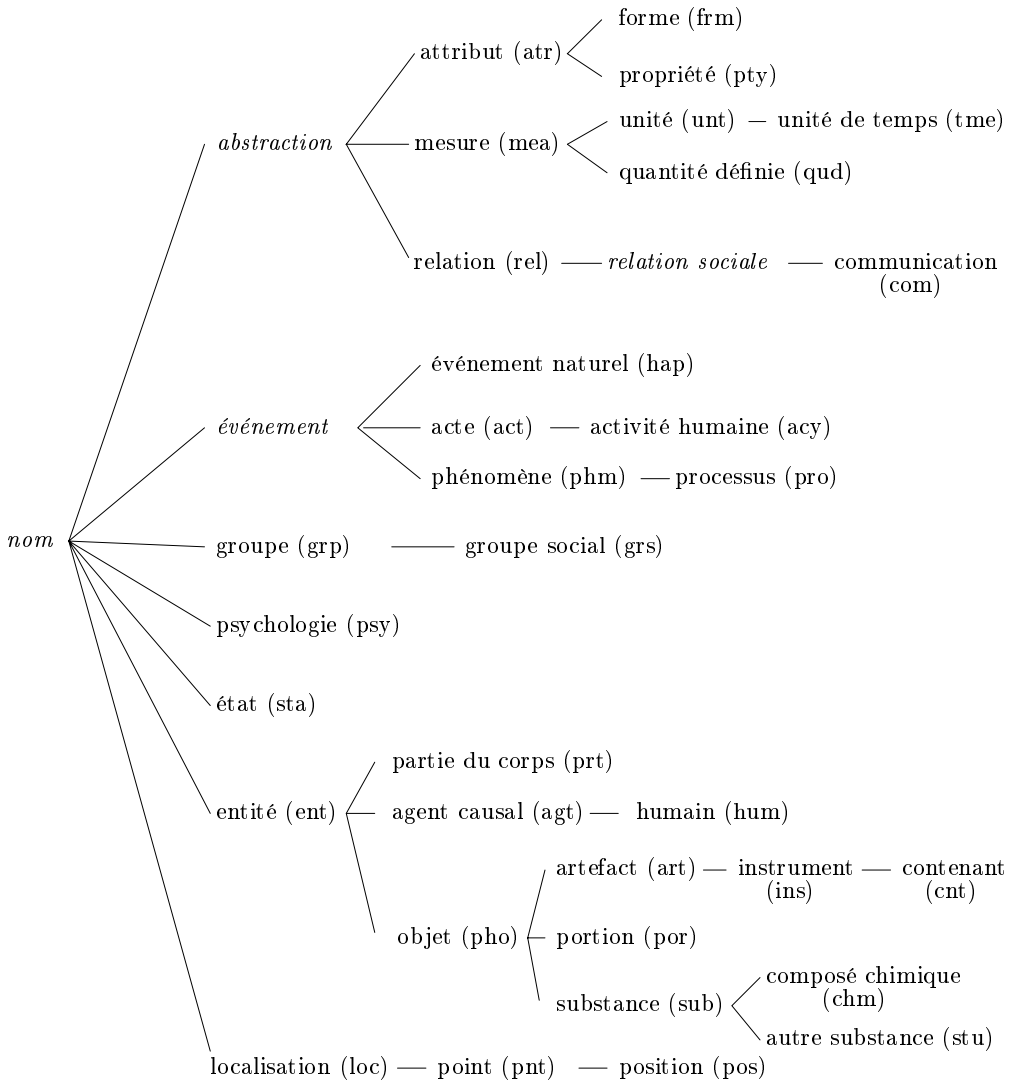

FIG. 4.1 – *Hiérarchie de classes pour l'étiquetage sémantique des noms*

En ce qui concerne les verbes, la classification de WordNet a été jugée inadaptée du fait d'un trop grand éparpillement des classes. Nous avons donc adopté une partition minimale en sept classes dans laquelle très peu de verbes sont ambigus (seulement 6 sur près de 570). Les autres catégories de mots du corpus (prépositions, pronoms...) ont aussi été organisées en classes et rangées dans un lexique ; là encore, comparativement aux noms, peu d'entrées pour ces catégories sont ambiguës.

Quelle que soit la catégorie (N, V...), le taux globalement faible d'ambiguïtés peut peut-être s'expliquer par le fait que le lexique construit est lié aux occurrences, c'est-à-dire que les étiquettes ont été choisies à l'aide des occurrences effectives des mots dans le corpus. Un mot, potentiellement ambigu dans le domaine étudié, mais pour lequel les occurrences constatées soulignent toutes le même sens, a reçu une étiquette unique dans le lexique.

L'étiquetage sémantique consiste alors à projeter sur chaque mot du corpus (déjà étiqueté catégoriellement) le contenu de l'entrée correspondante du lexique dont nous venons de décrire le mode de constitution. Les ambiguïtés sont ensuite résolues en utilisant, comme pour l'étiquetage catégoriel, un *étiqueteur* à chaînes de Markov cachées [13]. Comme mentionné ci-dessus, les ambiguïtés à résoudre sont principalement des problèmes de polysémie complémentaire, puisque les mots ont déjà subi une désambiguïsation catégorielle qui limite la polysémie contrastive. Une portion du texte de près de 6 000 mots a servi à mesurer la précision de notre étiquetage sémantique. Dans cet extrait, 7,78 % des mots étaient initialement ambigus, et l'étiquetage a permis de résoudre correctement 85 % de ces ambiguïtés.

Ces étiquetages catégoriel et sémantique du corpus nous permettent de disposer d'informations sur le contexte de couples N-V et de constituer les exemples nécessaires à l'apprentissage.

4.2 Méthode d'apprentissage

La mise au point de notre méthode d'apprentissage dans un cadre de PLI implique que nous fournissions un ensemble de paires N-V qualia (l'ensemble E^+ des exemples positifs) dans un contexte catégoriel et sémantique (c'est-à-dire des éléments des phrases contenant ces paires N-V dans notre corpus d'apprentissage), et un ensemble de paires N-V non qualia (ensemble E^-). L'une des difficultés majeures pour la production de règles généralisées à partir de ces contextes des paires qualia et d'informations de distance entre le N et le V dans les phrases concernées réside dans la quantité de données que l'algorithme de PLI doit gérer. Nous devons donc nous focaliser sur l'efficacité de l'étape d'apprentissage pour être certaine d'obtenir des clauses en un temps raisonnable. L'autre point important concerne le caractère expressif et significatif, sur le plan linguistique, des règles apprises. Nous ne voulons pas seulement obtenir des règles quelconques, efficaces pour extraire des couples qualia, mais des clauses qui aient linguistiquement du sens. La plupart des systèmes de PLI existants fournissent un moyen de traiter le problème de la forme des règles

13. Nous nous sommes inspirée, pour cet étiquetage sémantique, de principes décrits dans [RBB+99] et [BBRR00], qui présentent l'étiquetage d'un corpus du domaine médical.

que l'on souhaite obtenir, mais seuls certains d'entre eux permettent un contrôle complet de cette forme et de l'efficacité de la recherche de ces hypothèses. Pour cette double nécessité de contrôle, nous avons abandonné PROGOL, l'algorithme de PLI dû à Muggleton [Mug95] utilisé lors de nos deux premières expériences d'apprentissage (*cf.* introduction de ce chapitre) au profit d'ALEPH, l'implémentation de la PLI de Srinivasan[14], qui a déjà prouvé sa capacité à traiter des volumes importants d'informations dans de multiples domaines.

Dans cette section, nous décrivons la construction des exemples et des connaissances préalables pour ALEPH, puis nous expliquons comment, en définissant, en particulier, une relation de généralité entre les hypothèses bien adaptée à notre problème et un opérateur de raffinement permettant, parmi ces hypothèses, d'obtenir la meilleure d'entre elles, nous conduisons l'algorithme à produire efficacement des hypothèses bien formées et linguistiquement significatives.

4.2.1 Construction des exemples et des connaissances préalables

Étant donné un sous-ensemble de paires N-V de notre corpus[15], chacune des occurrences de ces paires est présentée en contexte à un expert qui les annote manuellement comme qualia ou non qualia. Chaque paire annotée qualia sert à constituer un exemple positif : un programme *Perl* ajoute une clause de la forme est_qualia(identificateur_N,identificateur_V). à l'ensemble E^+ et stocke dans les connaissances préalables (*background knowledge*) l'information qui décrit chaque mot de la phrase où apparaît cette paire, ainsi que la position de la paire et la distance entre ses éléments. Par exemple, pour une phrase de 5 mots dont les identificateurs sont m_1 ... m_5, et où la paire N-V est m_4-m_2, les clauses suivantes sont ajoutées :

```
tags(m_1,tag_catégoriel,tag_sémantique).
tags(m_2,tag_catégoriel,tag_sémantique).
pred(m_2,m_1).
tags(m_3,tag_catégoriel,tag_sémantique).
pred(m_3,m_2).
tags(m_4,tag_catégoriel,tag_sémantique).
pred(m_4,m_3).
tags(m_5,tag_catégoriel,tag_sémantique).
pred(m_5,m_4).
distances(m_4,m_2,distance en mots,distance en verbes).
```

où pred(x,y) indique que le mot y est situé juste avant le mot x dans la phrase, le prédicat tags/3 donne les étiquettes catégorielle et sémantique d'un mot, et distances/4 spécifie la distance[16] en nombre de mots et de verbes entre N et V dans la phrase[17].

Ainsi, l'exemple positif tiré de la phrase « *L'installation se compose : de deux atterrisseurs protégés par des carénages, fixés et articulés...* » est, par exemple,

14. http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/aleph/aleph_toc.html
15. *Cf.* [CSFB02] ou [CSBF01] pour la façon dont ces paires sont choisies.
16. Positive si V précède N, négative sinon, et décalée d'une unité pour différencier l'ordre dans le cas d'une distance nulle.
17. Certaines catégories de mots ne sont pas prises en considération dans le codage des exemples (les déterminants, certains adjectifs...).

simplement noté :

```
est_qualia(n609,v609).,
```

et les informations concernant ce couple N-V et son contexte sont stockées dans le *background knowledge* sous la forme :

```
tags(m609_8,tc_prep,ts_rman).
pred(n609,m609_8).
tags(m609_10,tc_wpunct,ts_virg).
suc(n609,m609_10).
tags(m609_6,tc_noun_pl,ts_art).
pred(v609,m609_6).
suc(v609,m609_8).
tags(n609,tc_noun_pl,ts_art).
tags(v609,tc_verb_adj,ts_acp).
distances(n609,v609,2,1).
```

suc indiquant le successeur d'un mot.

Les exemples négatifs sont construits de la même manière à partir des paires annotées non qualia. D'autres clauses décrivant les relations entre les étiquettes sémantiques ou catégorielles sont aussi ajoutées au *background knowledge* d'ALEPH. Ces relations indiquent par exemple que l'étiquette tc_verb_pl correspond à un verbe conjugué au pluriel (conjugue_pluriel), qui peut être considéré comme un verbe conjugué (conjugue) ou simplement un verbe (verbe). Voici un exemple de ces littéraux décrivant les mots d'un point de vue linguistique :

```
verbe( M ) :- conjugue( M ).
verbe( M ) :- infinitif( M ).
...
conjugue( M ) :- conjugue_pluriel( M ).
conjugue( M ) :- conjugue_singulier( M ).
conjugue_pluriel( M ) :- tagcat( M, tc_verb_pl ).
...
```

ALEPH tente alors de gérer au mieux cette grande quantité d'informations (les exemples positifs et négatifs et le *background knowledge*) et de découvrir des règles expliquant la plupart des E^+ et rejetant la plupart des E^-. Pour inférer ces règles, il utilise les exemples pour générer et tester diverses hypothèses, et conserve celles qui semblent les plus pertinentes par rapport à ce que nous voulons apprendre. Nous présentons maintenant comment nous avons obtenu de manière efficace des règles linguistiquement motivées, en précisant un langage d'hypothèses et en définissant une relation de généralité entre les hypothèses et un opérateur de raffinement permettant la production de règles bien formées couvrant le maximum E^+ et le minimum d'E^-.

4.2.2 Production efficace d'hypothèses bien formées

Afin de faciliter la compréhension des différents aspects de l'algorithme d'apprentissage (opérateur de raffinement, élagage...) sur lesquels nous avons travaillé, nous présentons, dans un premier temps le fonctionnement général d'un algorithme de PLI tel qu'ALEPH.

L'algorithme utilisé par ALEPH (et de nombreux autres algorithmes de PLI) peut se découper en quatre étapes :

1. sélectionner un E^+ ;
2. construire la clause la plus spécifique à partir de cet exemple. Cette clause (appelée *bottom clause*) est en fait l'hypothèse la moins générale qui couvre l'exemple sélectionné. Le processus de construction de la *bottom clause* (appelé parfois saturation) est détaillé dans [Mug95] ;
3. rechercher la meilleure hypothèse. Cette hypothèse doit être plus générale que la *bottom clause* ;
4. ôter les exemples couverts par l'hypothèse retenue et retourner au point 1.

L'opérateur de raffinement intervient lors de la troisième étape. Il permet de rechercher dans l'ensemble des hypothèses possibles la clause qui, relativement à une certaine fonction de score, se révèle la meilleure. Cet ensemble de clauses est un espace ordonné par une notion de généralité. Au sommet de cet espace – qui peut être un treillis – se trouve la clause la plus générale (dans notre cas, c'est `est_qualia(A,B)` : tout couple N-V est qualia) et la *bottom clause* est, quant à elle, la borne inférieure. La recherche de l'hypothèse maximisant la fonction de score se fait en parcourant l'espace entre ces deux clauses (cf. figure 4.2).

Le parcours se fait généralement de haut en bas, c'est-à-dire du plus général au plus spécifique. Ainsi, à partir d'une hypothèse se révélant trop générale (couvrant trop d'exemples négatifs) est générée[18] une autre hypothèse plus spécifique. Le score de cette dernière est alors évalué et le processus est réitéré jusqu'à être certain qu'il n'y ait pas dans l'espace une meilleure hypothèse. Il n'est généralement pas nécessaire de tester exhaustivement toutes les clauses pour s'assurer de cette dernière condition ; il est en effet souvent possible d'élaguer l'espace d'hypothèses au cours de la recherche, et cet élagage permet une amélioration importante de la rapidité de l'algorithme. Un bon opérateur doit donc permettre autant que faire se peut d'élaguer l'espace de la manière la plus achevée possible. Il cherche également à respecter, si possible, certaines propriétés [vdL95], dont la complétude (toutes les hypothèses peuvent être atteintes), la non redondance (il n'y a qu'une façon d'accéder à une hypothèse)... L'écriture d'un opérateur de raffinement est donc un point crucial d'efficacité et d'expressivité pour tout problème de PLI.

Pour mettre au point notre méthode d'apprentissage de couples N-V qualia, nous avons donc dû préciser un langage pour les hypothèses inférées ; nous avons également défini un ordre de généralité adapté à notre problème pour organiser l'espace d'hypothèses qui se révèle être un treillis, et proposé un opérateur de raffinement permettant de parcourir de façon « intelligente » ce treillis, en se basant sur une fonction de score que nous allons préciser, et en évitant de parcourir des branches inutiles grâce à un élagage dont nous allons expliquer les principes. La fin de cette section est dédiée à l'explicitation de ces divers éléments décrits en détail dans [CSFB02].

18. C'est ce processus qui a conduit Shapiro à appeler ce système de génération d'hypothèses *opérateur de raffinement* [Sha81].

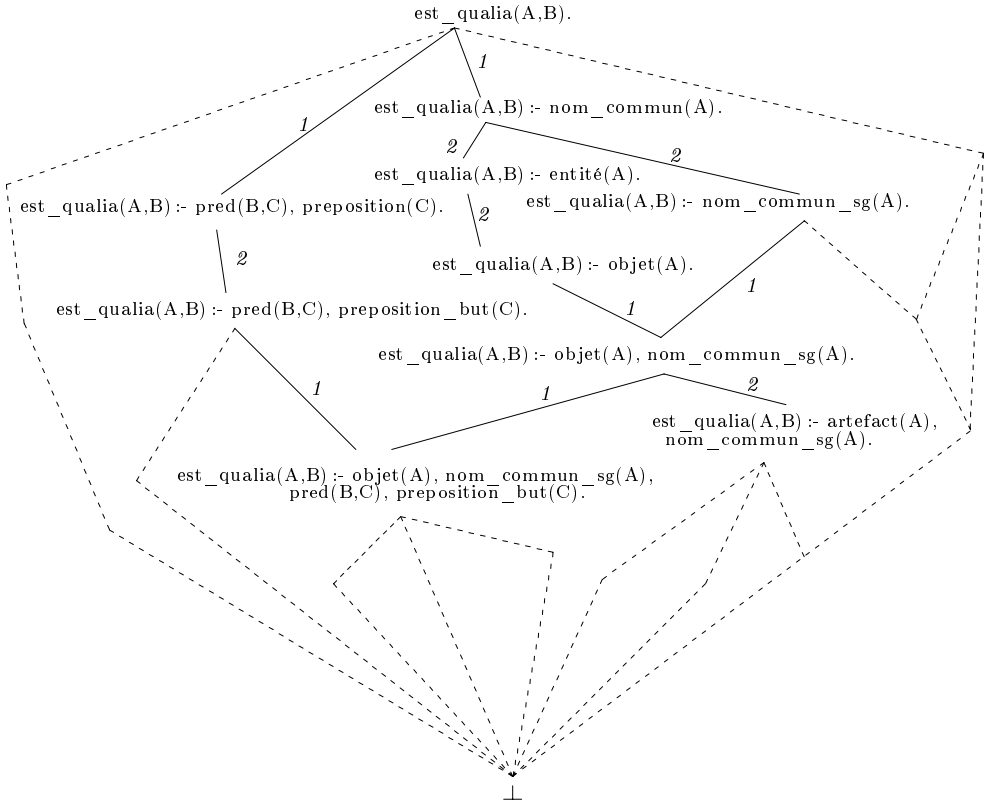

FIG. 4.2 – *Exemple de treillis de recherche des hypothèses*

Langage d'hypothèses

La première phase de développement d'un système d'apprentissage efficace et produisant des règles dotées d'un intérêt consiste donc à déterminer un langage d'hypothèses. En effet, en PLI, on cherche à inférer une hypothèse H vérifiant :

$\forall e^+ \in E^+ : B \cup H \models e^+$ (complétude)

$\forall e^- \in E^- : B \cup H \not\models e^-$ (correction)

où B est le *background knowledge*. Une telle hypothèse pouvant *a priori* être cherchée dans l'espace de toutes les clauses de Horn, de nombreux biais évitant d'examiner tout cet espace (*cf.* [NRA+96]) ont été proposés, dont le biais sur le langage d'hypothèses qui spécifie des contraintes syntaxiques sur les hypothèses à trouver. Pour nous, une hypothèse bien formée est définie comme une clause qui donne des informations (sémantiques ou catégorielles) sur les mots (le N, le V ou des mots de leur contexte) et des informations sur les positions respectives du N et du V dans la

phrase. Par exemple est_qualia(A,B)[19] :- artefact(A), pred(B,C), suc(A,C), auxiliaire(C). – qui signifie qu'une paire N-V est qualia si N est un artefact, V est précédé d'un auxiliaire et N est suivi par ce même verbe – est une hypothèse bien formée. Nous devons donc indiquer dans le paramétrage d'ALEPH que les prédicats artefact/1, pred/2, suc/2, auxiliaire/1... peuvent être utilisés pour construire une hypothèse. Une autre contrainte sur le langage d'hypothèses est qu'il ne peut y avoir qu'au plus une information catégorielle et une information sémantique pour un mot donné. Ceci veut dire que l'hypothèse est_qualia(A,B) :- pred(B,C), participe(C), participe_passé(C). n'est pas légale car le mot représenté par C est caractérisé par deux informations catégorielles. Cette information redondante sur un mot est superflue car les informations catégorielles et sémantiques sont organisées de façon hiérarchique, et un des littéraux est donc plus spécifique que les autres et décrit le mot plus précisément ; notre opérateur de raffinement doit donc gérer ce problème. Plusieurs autres prédicats, en particulier ceux traitant de distance entre le N et le V et de leurs positions relatives, sont utilisés dans le langage d'hypothèses. Plus de 100 prédicats différents peuvent ainsi apparaître dans une hypothèse.

Généralité : θ_{NV}-subsomption

Même avec ce biais de langage, l'espace de recherche demeure donc immense. Les hypothèses peuvent heureusement être organisées selon une relation de généralité qui permet à l'algorithme de parcourir « intelligemment » l'espace des solutions. Beaucoup de systèmes de PLI utilisent pour tel ordre partiel la θ-subsomption [Plo70], définie par :

Définition 1 (θ-subsomption) : *Une clause C_1 θ-subsume une clause C_2 ($C_1 \succeq_\theta C_2$) si et seulement si (ssi) il existe une substitution θ telle que $C_1\theta \subseteq C_2$ (en considérant les clauses comme des ensembles de littéraux).*

Cependant cette notion de généralité n'est pas adaptée à notre cas. En effet, soit $H_1 \equiv$ est_qualia(X_1,Z_1) :- suc(X_1,Y_1), pred(Z_1,W_1), verbe(Y_1), verbe(W_1). et $H_2 \equiv$ est_qualia(X_2,Z_2) :- suc(X_2,Y_2), pred(Z_2,Y_2), verbe(Y_2).. $H_1 \succeq_\theta H_2$ avec $\theta = [X_1/X_2, Y_1/Y_2, Z_1/Z_2, W_1/Y_2]$; puisque dans notre application, les variables représentent des mots, ceci signifie que la θ-subsomption permet de considérer un même mot comme deux mots différents dans une clause, comme c'est le cas pour le mot Y_1/W_1 dans H_1, ce qui, pour notre problème, n'est pas jugé pertinent. Nous avons donc basé notre propre notion de généralité sur une forme plus faible que la θ-subsomption : la θ-subsomption sous identité d'objet (notée par la suite θ_{OI}-subsomption) [ELMS96] définie par[20] :

Définition 2 (θ_{OI}-subsomption) : *Une clause C_1 θ_{OI}-subsume une clause C_2 ($C_1 \succeq_{OI} C_2$) ssi il existe une substitution θ telle que $C_1\theta \subseteq C_2$ et θ est injective (c'est-à-dire que θ n'unifie pas des variables de C_1).*

Cette forme de généralité est gérée par ALEPH en générant des hypothèses contenant des ensembles d'inégalités indiquant que des variables ayant des noms différents

19. ALEPH nomme automatiquement les variables ; par conséquent, dans les exemples cités, A correspond généralement au N et B au V des couples N-V.

20. D'après [BS99].

ne peuvent être unifiées. Par exemple, H_1 est représentée dans ALEPH par :

```
est_qualia(X,Z):- suc(X,Y), pred(Z,W), verbe(Y), verbe(W), X≠Z, X≠Y, Z≠Y,
X≠W, Y≠W, Z≠W.
```

Par la suite, pour faciliter la lecture, nous n'écrivons pas ces ensembles d'inégalités et considérons donc que deux variables de noms différents sont distinctes.

Par rapport à la θ_{OI}-subsomption, la θ_{NV}-subsomption que nous avons définie prend en compte l'organisation hiérarchique arborescente de nos informations catégorielles et sémantiques[21]. Ainsi, nous voulons que l'hypothèse est_qualia(A,B):- objet(A). soit considérée comme plus générale que est_qualia(A,B):- artefact(A). (*cf.* figure 4.1). De plus, nous souhaitons éviter la production de clauses contenant des variables non contraintes. Par exemple, l'hypothèse est_qualia(A,B):- infinitif(B), pred(A,C). pourrait être exprimée par est_qualia(A,B):- infinitif(B). car pred(A,C) n'apporte aucune information linguistiquement intéressante. Des idées assez similaires ont par exemple été exprimées dans la contrainte de connexion par Quinlan [Qui90], mais dans notre cas, chaque variable doit non seulement être reliée à une variable de la tête de clause par un *chemin* de variables (grâce à pred/2 et suc/2), mais doit en plus être « utilisée » quelque part dans le corps de l'hypothèse. Une hypothèse respectant toutes ces conditions est dite bien formée par rapport à notre tâche d'apprentissage.

Définition 3 (θ_{NV}-subsomption) : *Par conséquent, $C \succeq_{NV} D$ s'il existe une substitution injective θ et une fonction f_D telles que $f_D(C)\theta \subseteq D$*[22] *où f_D est telle que $\forall l \in C,\ B$*[23]*$, f_D(l) \models l$.*

Intuitivement ceci signifie qu'une clause D peut être plus spécifique que C si

1 – D a des littéraux supplémentaires par rapport à C;

2 – D contient des littéraux plus spécifiques (par rapport aux hiérarchies d'informations catégorielle et sémantique) sur les mêmes variables que C.

Grâce à la représentation de nos exemples et au *background knowledge* utilisé, tous les littéraux pouvant apparaître dans une clause sont déterministes ; de telles hypothèses sont dites clauses déterministes. L'ordre partiel de la θ_{NV}-subsomption implique que l'espace d'hypothèses est un treillis (*cf.* démonstration dans [CSFB02]). En haut se trouve la clause la plus générale ($\top$) et en bas, la plus spécifique (appelée *bottom* et notée $\bot$). Comme nous l'avons déjà mentionné, dans notre cas, $\top$ est la clause est_qualia(A,B). affirmant que toutes les paires N-V sont qualia, et $\bot$ est une clause sans constante contenant tous les littéraux qui peuvent décrire l'exemple à généraliser, moins quelques littéraux superflus (des littéraux fournissant des informations plus générales sur un mot que d'autres littéraux de $\bot$). La figure 4.2 présente un exemple simple de notre treillis ; les chiffres sur les arcs font référence aux première et seconde conditions de la définition de la θ_{NV}-subsomption.

21. Suivant en cela des idées développées dans le cadre de la subsomption généralisée [Bun88].
22. $f_D(\{l_1,l_2,...,l_m\})$ signifie $\{f_D(l_1),f_D(l_2),...,f_D(l_m)\}$.
23. *Background knowledge.*

Opérateur de raffinement et élagage

Pour parcourir ce treillis de manière efficace, nous avons défini un opérateur de raffinement non redondant (ρ_{nv}) qui a pour but de parcourir le treillis d'hypothèses à la recherche de la meilleure, étant donnée une fonction de score. Conformément à la définition de la θ_{NV}-subsomption (*cf.* définition 3), ρ_{nv} est formé de 12 sous-opérateurs ayant pour fonction d'ajouter une variable à une clause, ou d'ajouter (ou affiner) de l'information portant sur les variables d'une clause[24]. La recherche d'hypothèse doit par ailleurs être faite en évitant de parcourir des zones de l'espace de recherche où aucune « bonne » solution n'est susceptible d'être trouvée ; l'opérateur de raffinement doit donc ne pas tenter de raffiner une hypothèse dont on sait qu'aucun descendant ne sera pertinent. Pour définir un élagage sûr, nous nous sommes basée sur la notion de propriété privée de Torre et Rouveirol [TR97c, TR97a, TR97b][25].

Définition 4 (Propriété privée) : *Une propriété P est dite privée étant donné un opérateur de raffinement ρ de l'espace de recherche S ssi :*

$$\forall H,H' \in S : \forall(H' \in \rho^*(H) \wedge \overline{P(H)} \Rightarrow \overline{P(H')})$$

où $\overline{X}$ est la négation de X, $\rho^*(H)$ est l'ensemble des hypothèses atteignables par raffinements successifs à partir de H par ρ, et $\forall F$, pour une formule F, représente la fermeture universelle de F, c'est-à-dire la formule close obtenue en ajoutant un quantificateur universel à chaque variable ayant une occurrence libre dans F.

Parmi les nombreuses propriétés privées possibles, nous utilisons la longueur de la clause proposée dans [TR97c], qui limite à k littéraux la longueur d'une hypothèse. Cette propriété est privée pour ρ_{nv} car notre opérateur consiste (*cf.* définition 3) soit à ajouter des littéraux (longueur supérieure), soit à affiner des littéraux (longueur identique). Nous prenons aussi en compte le nombre minimum d'E^+ couverts, en évitant de parcourir des raffinements d'une hypothèse ne couvrant pas un nombre suffisant d'E^+, puisque l'on sait que ce nombre diminue lors de la spécialisation. Enfin, l'élagage est également souvent basé en PLI sur la fonction de score qui permet de définir la meilleure hypothèse pour la tâche d'apprentissage. Nous avons choisi la fonction s définie pour une hypothèse H par : $s(H) = (P - N,|H|)$ où P (respectivement N) est le nombre d'exemples positifs (respectivement négatifs) couverts par H et $|H|$ est la longueur de H, soit le nombre de littéraux qu'elle contient. H_1 est dite meilleure hypothèse que H_2 (avec $s(H_1) = (P_1 - N_1,|H_1|)$ et $s(H_2) = (P_2 - N_2,|H_2|)$) ssi $P_1 - N_1 > P_2 - N_2$ ou $P_1 - N_1 = P_2 - N_2 \wedge |H_1| < |H_2|$. Cependant, $P - N$ n'étant pas monotone, nous ne pouvons rien affirmer sur les scores des raffinements d'une hypothèse donnée H qui satisfait un certain critère de score tel que $s(H) < k$, où k peut, par exemple, être le meilleur score trouvé jusqu'ici lors de la recherche. La propriété privée liée à cette fonction de score que nous utilisons pour l'élagage est donc plus faible : $s_{opt}(H) \geq S_{meilleur}$, où $S_{meilleur}$ est la plus grande différence $P - N$ trouvée lors de la recherche et $s_{opt}(H) = P_{courant} - N_{\perp}$.

24. Pour information, cet opérateur est *parfait* [BS99], c'est-à-dire minimal et optimal (fini, non redondant et faiblement complet) ; *cf.* [CSFB02] pour plus de détails.

25. Nous tenons à remercier Céline Rouveirol pour ses suggestions et pour les discussions enrichissantes que nous avons eues avec elles.

$P_{courant}$ est le nombre d'exemples positifs couverts par l'hypothèse courante, $N_{\perp}$ est le nombre d'exemples négatifs couverts par $\perp$ (évaluée à sa construction). Cette propriété est bien privée, car $\forall H,H' \in S : \forall S_{meilleur} \in \mathbb{N} : \ (H' \in \rho^*(H) \wedge s_{opt}(H) < S_{meilleur} \Rightarrow s_{opt}(H') < S_{meilleur})$ car P diminue lors de la recherche et $N_{\perp}$ est constant.

Le parcours structuré de notre espace de recherche nous permet des gains importants d'efficacité et ramène les temps d'apprentissage, en prenant en compte l'intégralité des étiquetages catégoriel et sémantique, à ceux obtenus lors des deux premières expériences mentionnées dans l'introduction de ce chapitre (soit quelques heures sur un PC 966MHz sous Linux). Le processus d'apprentissage produit deux types de sortie : quelques exemples positifs non généralisés, et un ensemble G de clauses généralisées sur lequel nous allons maintenant nous focaliser dans le cadre de la prise en compte de toutes les informations contextuelles.

4.3 Apprentissage basé sur les informations catégorielles et sémantiques

Nous avons fourni à ALEPH 3099 E^+ et 3176 E^- construits selon la méthode et le format précisés en section 4.2.1, et avons effectué un apprentissage en prenant en considération les informations catégorielles et sémantiques associées aux mots. L'objectif de cette partie est de présenter les règles obtenues et les résultats de l'application de ces clauses sur le corpus MATRA-CCR pour extraire des couples N-V qualia. Plusieurs étapes d'évaluation et de validation de la méthode d'apprentissage sont cependant nécessaires pour garantir que les règles inférées soient pertinentes pour la tâche d'extraction de ces paires nomino-verbales. Nous décrivons donc dans un premier temps l'évaluation théorique de l'apprentissage et du choix de ses paramètres. Nous présentons ensuite les règles produites et une évaluation empirique de leurs performances pour acquérir des couples qualia. Nous comparons les résultats de notre système d'extraction à des techniques statistiques ou syntaxiques simples, pour pointer ses spécificités, avantages et inconvénients, et terminons par une discussion sur la validité linguistique des clauses générales apprises.

4.3.1 Validation théorique de l'apprentissage

Lorsqu'ALEPH reçoit en entrée les exemples et les connaissances préalables, l'apprentissage produit un ensemble G de clauses généralisées. Cependant, avant d'utiliser ces règles pour extraire des couples qualia, nous devons vérifier qu'elles ont été correctement apprises, c'est-à-dire que le paramétrage de l'algorithme a été bien réalisé, que les exemples étaient suffisamment représentatifs et nombreux...

Un de ces paramètres fondamentaux, en particulier quand on manipule des données issues d'un traitement automatique de corpus, est le bruit, nombre d'E^- que les règles apprises ont le droit de couvrir. La prise en compte du bruit peut permettre d'obtenir des règles plus générales, en acceptant d'expliquer quelques E^- qui peuvent être, par exemple, liés à des erreurs d'étiquetage. Nous avons testé différentes valeurs de ce paramètre et évalué leurs effets. Les résultats des expériences

successives (en particulier le rappel et la précision des règles en termes de couverture des exemples) sont comparés à l'aide d'une mesure unique, le coefficient de Pearson (coefficient Phi) qui s'exprime de la façon suivante :

$$Pearson = \frac{(TP * TN) - (FP * FN)}{\sqrt{PrP * PrN * AP * AN}}$$

où A = *actual* (réel), Pr = *predicated* (prédit), P = *positive*, N = *negative*, T = *true*, F = *false* ; une valeur proche de 1 indique un bon apprentissage.

Afin d'estimer de manière précise les diverses caractéristiques de cette étape d'apprentissage et, en particulier, de garantir l'indépendance des résultats obtenus par rapport à l'ordre de présentation des exemples à l'algorithme, nous avons, pour chacun de ces tests, effectué une validation croisée (*10-fold cross-validation* [Koh95]). Nous avons ainsi décomposé l'ensemble des 3099 exemples positifs et des 3176 négatifs en 10 sous-ensembles. Chaque sous-ensemble est à tour de rôle utilisé comme ensemble de test, alors que les 9 autres forment l'ensemble d'apprentissage de l'algorithme de PLI. Dix apprentissages sont donc réalisés sur les ensembles d'apprentissage et testés sur l'ensemble de test correspondant. Les résultats – moyenne des temps de calcul, de la précision, du rappel et du coefficient de Pearson, ainsi que les écarts-types – obtenus avec le taux de bruit optimal (*i.e.* celui qui maximise le coefficient de Pearson) sont résumés dans le tableau 4.1.

	Temps (secondes)	Précision (%)	Rappel (%)	Coefficient de Pearson
Moyenne	10285	81.3	89.0	0.693
Écart-type	1440	2.8	2.4	0.047

TAB. 4.1 – *Résultats de la validation croisée*

Un apprentissage final est alors mené en prenant en compte l'intégralité des E^+ et des E^- comme ensemble d'apprentissage.

4.3.2 Résultats et validation empirique

Avec ce paramétrage optimal, ALEPH produit les 9 règles générales suivantes :

(1) est_qualia(A,B) :- precede(B,A), proche_verbe(A,B), infinitif(B), verbe_action(B).
(2) est_qualia(A,B) :- contigu(A,B).
(3) est_qualia(A,B) :- precede(B,A), proche_mot(A,B), proche_verbe(A,B), suc(B,C), preposition(C).
(4) est_qualia(A,B) :- proche_mot(A,B), pred(A,C), vide(C).
(5) est_qualia(A,B) :- precede(B,A), suc(B,C), pred(A,D), ponctuation(D), nom_commun_sg(A), deux_points(C).
(6) est_qualia(A,B) :- proche_mot(A,B), suc(B,C), suc(C,D), verbe_action(D).
(7) est_qualia(A,B) :- precede(A,B), proche_mot(A,B), pred(A,C), ponctuation(C).

(8) est_qualia(A,B):- proche_verbe(A,B), pred(B,C), pred(C,D), pred(D,E), preposition(E), pred(A,F), vide(F).
(9) est_qualia(A,B):- precede(A,B), proche_verbe(A,B), pred(A,C), conjonction_subordination(C).

où precede(X,Y) signifie que X précède Y, et pred(X,Y) (respectivement suc(X,Y)) que Y est le prédécesseur (respectivement successeur) immédiat de X ; proche_mot(X,Y) signifie que X et Y sont séparés par au moins un mot et au plus deux, et proche_verbe(X,Y) qu'il n'y a pas de verbe entre X et Y.

La première clause indique donc qu'une paire N-V est qualia si le V précède le N (sans aucun autre verbe entre eux) et que ce V est un verbe d'action à l'infinitif, comme dans la phrase « *enclencher le disjoncteur* », où enclencher joue le rôle télique de disjoncteur. Nous revenons plus en détail sur une discussion linguistique de ces règles en section 4.3.4.

La validation empirique de notre méthode d'apprentissage est réalisée en appliquant ces 9 clauses sur le corpus MATRA-CCR et en étudiant la pertinence de leurs décisions concernant le caractère qualia ou non de couples N-V, en les comparant à celles d'experts du LG. Plus précisément, nous avons effectué cette validation sur un sous-ensemble de 32 000 mots du corpus et, pour des raisons de faisabilité[26], nous nous sommes focalisée sur 7 noms jugés significatifs dans le domaine de ce corpus : *vis, écrou, porte, voyant, prise, capot, bouchon*[27].

Dans un premier temps, un programme *Perl* repère sur ce sous-corpus chacune des occurrences des paires N-V incluant un de ces 7 noms et un verbe quelconque de la même phrase. Quatre experts étiquettent manuellement chaque paire comme qualia ou non, les divergences étant discutées jusqu'à accord complet. Ceci permet d'atteindre un total de 66 paires N-V qualia parmi les 286 contenant un des 7 noms retenus.

Les 9 règles sont donc ensuite appliquées sur le sous-corpus, et les décisions prises par elles sur le caractère qualia ou non des 286 paires étudiées sont comparées à celles des experts. Il nous faut toutefois choisir le nombre d'occurrences x d'une même paire qui doivent être reconnues par les règles apprises pour que cette paire soit considérée comme qualia par elles. Une valeur élevée de x favorise la précision de l'extraction, alors qu'une valeur faible favorise le rappel. Pour fixer ce seuil, nous choisissons la valeur qui maximise le coefficient de Pearson, qui dans notre cas est 1. Par conséquent, une paire N-V est considérée comme qualia si (au moins) une de ses occurrences est couverte par une des 9 clauses. Le tableau 4.2 résume les résultats que nous obtenons sur notre ensemble de test empirique ; nous y indiquons également la valeur de la F-mesure[28], moyenne harmonique pondérée des taux de rappel et précision, communément utilisée en RI pour comparer les performances de systèmes.

Nous obtenons un très bon rappel et une bonne précision. Les 9 règles apprises

26. Il est en effet impossible d'examiner les décisions prises par les règles pour l'intégralité des couples N-V d'un tel sous-corpus, même de taille réduite.

27. Ces noms n'ont pas participé à la constitution des exemples d'apprentissage.

28. $F = \frac{PR}{(1-\alpha)P+\alpha R}\ 0 \leq \alpha \leq 1$, où R est le taux de rappel et P celui de précision. La valeur la plus communément considérée pour α est 0,5, soit $F = \frac{2PR}{P+R}$.

	Rappel (%)	Précision (%)	F-mesure	Coefficient de Pearson
Système de PLI	92.4	62.9	0.748	0.677

TAB. 4.2 – *Résultats empiriques de la méthode d'apprentissage de type PLI*

semblent donc décrire correctement le concept de rôle qualia et nous pouvons appliquer ce système d'extraction de paires qualia à l'ensemble du corpus. Une discussion détaillée des types de paires correctement retrouvées, oubliées ou incorrectement détectées est présentée en section 4.3.4.

4.3.3 Comparaison avec des approches statistiques et syntaxiques

Comme nous l'avons mentionné en introduction de ce chapitre, plusieurs méthodes statistiques ou à base de patrons morpho-syntaxiques ont déjà été utilisées pour extraire des cooccurrences. Nous avons comparé les performances de notre système fondé sur la PLI pour l'acquisition de couples N-V qualia avec des techniques statistiques simples et une méthode syntaxique manuelle. Ce souhait de comparaison avec deux types de méthodes habituelles d'acquisition d'éléments en relation sémantique au sein de corpus est motivé par le fait que notre approche peut être vue comme se situant entre ces deux optiques. Comme les méthodes statistiques, nous cherchons à inférer des couples N-V dont les constituant sont liés en traitant automatiquement des corpus ; cependant, contrairement à elles, ce ne sont pas uniquement des paires dont les éléments entretiennent un lien statistiquement significatif que nous voulons apprendre, mais des couples dont les composants sont reliés par des liens très spécifiques, définis par une théorie linguistique ; de plus, nous voulons non seulement obtenir ces couples, mais apprendre automatiquement des règles linguistiquement motivées expliquant ce qui caractérise les liens dans ces couples. Les approches linguistiques « classiques » utilisent la reconnaissance de patrons morpho-syntaxiques, fonctionnels... pour recenser des éléments exprimant une relation donnée ; un lien syntaxique peut donc être vu comme marqueur d'une relation – par exemple sémantique – particulière ; toutefois, pour nous, les patrons marqueurs des liens qualia ne sont pas connus *a priori*, et nous cherchons à les inférer automatiquement, pour chaque corpus choisi. Nous voulons donc étudier ce que chacune des méthodes apporte, et avons donc appliqué au même jeu de test que pour notre expérience des techniques statistiques et une méthode syntaxique manuelle ; nous présentons ici les résultats obtenus et les spécificités de chacun de ces trois types d'approches.

Méthodes statistiques

Nous avons choisi 10 mesures statistiques, connues et éprouvées dans de nombreux domaines, pour extraire des cooccurrences de noms et verbes et tester le

caractère qualia ou non des paires obtenues. Ces cooccurrences sont repérées au sein de phrases, en considérant les lemmes des mots.

À chaque paire N-V (N_j,V_i), nous associons un tableau de contingence tel le tableau 4.3, dans lequel a est le nombre d'occurrences de la paire N-V, b celui de paires N-V où le nom est N_j mais le verbe n'est pas V_i, c celui de paires N-V où le verbe est V_i mais le nom n'est pas N_j, et d celui des paires N-V où le nom n'est pas N_j et le verbe n'est pas V_i. Nous nommons S la somme a + b + c + d.

	V_i	$V_{i'}$, $i' \neq i$
N_j	a	b
$N_{j'}$, $j' \neq j$	c	d

TAB. 4.3 – *Tableau de contingence de la paire N-V (N_j,V_i)*

Les 10 critères d'association que nous avons retenus sont alors exprimés par [Dai94] :

- coefficient de Kulczinsky : $Kul = \frac{a}{2}\left(\frac{1}{a+b} + \frac{1}{a+c}\right)$
- coefficient d'Ochiai : $Ochiai = \frac{a}{\sqrt{(a+b)(a+c)}}$
- score d'information mutuelle : $IM = log_2 \frac{a}{(a+b)(a+c)}$
- score d'information mutuelle au cube [Dai94] : $IM^3 = log_2 \frac{a^3}{(a+b)(a+c)}$
- coefficient de McConnoughy : $MC = \frac{a^2-bc}{(a+b)(a+c)}$
- test d'association du χ^2 : $\chi^2 = \frac{\left(a-\frac{(a+b)(a+c)}{S}\right)^2}{\frac{(a+b)(a+c)}{S}} + \frac{\left(b-\frac{(a+b)(b+d)}{S}\right)^2}{\frac{(a+b)(b+d)}{S}} + \frac{\left(c-\frac{(c+d)(a+c)}{S}\right)^2}{\frac{(c+d)(a+c)}{S}} + \frac{\left(d-\frac{(c+d)(b+d)}{S}\right)^2}{\frac{(c+d)(b+d)}{S}}$
- coefficient de vraisemblance (loglike) [Dun93] : $loglike = a\log a + b\log b + c\log c + d\log d - (a+b)\log(a+b) - (a+c)\log(a+c) - (b+d)\log(b+d) - (c+d)\log(c+d) + S\log S$
- coefficient de proximité simple : $SMC = \frac{a+d}{S}$
- coefficient de Yule : $Yule = \frac{ad-bc}{ad+bc}$
- coefficient du Φ^2 [CG91] : $\Phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$

Ces mesures statistiques sont évaluées sur le même jeu de test que précédemment. Pour chacune d'entre elles, nous déterminons la valeur du coefficient d'association indiquant un seuil entre les cooccurrences retenues ou non, en prenant celle qui maximise le coefficient de Pearson, conformément à ce qui a été fait pour notre technique à base de PLI. Le tableau 4.4 indique les résultats optimaux ainsi obtenus.

Seules quelques mesures statistiques donnent des résultats permettant de les utiliser dans une tâche d'extraction de paires qualia, sans toutefois atteindre la qualité globale des scores obtenus par notre méthode d'apprentissage de PLI. Les différences entre notre système et ces techniques statistiques s'expliquent bien sûr par le fait que ces deux approches ne manipulent pas la même connaissance : les modèles

	Rappel (%)	Précision (%)	F-mesure	Coefficient de Pearson
Kul	36.4	70.6	0.48	0.414
$Ochiai$	42.4	82.4	0.56	0.517
IM	51.5	40	0.45	0.261
IM^3	36.4	92.3	0.522	0.52
MC	36.4	70.6	0.48	0.414
χ^2	37.9	78.1	0.51	0.464
$loglike$	42.4	80	0.554	0.505
SMC	100	25.3	0.385	0.17
$Yule$	53	41.2	0.464	0.279
Φ^2	37.9	78.1	0.51	0.464

TAB. 4.4 – *Résultats des méthodes statistiques*

statistiques n'utilisent que la cooccurrence de lemmes, alors que notre système tire parti des étiquettes catégorielles et sémantiques et requiert des exemples positifs et négatifs. Cette comparaison est toutefois intéressante pour argumenter sur l'équilibre entre le choix d'une méthode supervisée ou non et la qualité des performances résultantes.

Liens syntaxiques

Les résultats de notre technique sont aussi comparés avec ceux obtenus par une méthode reposant sur une analyse syntaxique manuelle du corpus. En fait, plus que de comparaison de performances, notre objectif est surtout ici de caractériser plus finement le lien qualia par rapport au lien syntaxique. L'idée sous-jacente est la suivante : une paire N-V dont les constituants sont fréquemment en relation syntaxique signale peut-être un lien sémantique entre ces éléments, potentiellement de type qualia. Nous avons donc annoté syntaxiquement, par de simples liens sujet, objet et modifieur, les phrases où apparaissent les 286 paires N-V de notre validation empirique. Une paire N-V est alors considérée comme qualia si un certain nombre de ses occurrences sont détectées comme étant syntaxiquement liées. Ce seuil est là aussi choisi pour maximiser le coefficient de Pearson, et la valeur optimale trouvée est 1. Le table 4.5 rassemble les performances d'un tel système sur notre ensemble test.

	Rappel (%)	Précision (%)	F-mesure	Coefficient de Pearson
Lien syntaxique	86.4	79.2	0.826	0.772

TAB. 4.5 – *Résultats de la méthode syntaxique manuelle*

Si le taux de rappel est légèrement plus faible que celui de notre système, le

taux de précision est en revanche plus élevé. Le taux de rappel inférieur à 100% tend à prouver qu'un lien qualia est plus qu'un simple lien syntaxique sujet, objet ou modifieur. Parmi les 13,6% de paires N-V qualia dont les constituants ne sont pas liés par un lien syntaxique, on trouve essentiellement des couples apparaissant dans des phrases contenant des ellipses ou utilisant certains signes de ponctuation, comme dans les exemples « éteindre le voyant ; allumer » (couple qualia *voyant - allumer* non syntaxiquement lié), « poser l'ensemble : rondelle, vis et serrer au couple » (couple qualia *poser - vis* non syntaxiquement lié), « enlever le bouchon : nettoyer » (couple qualia *bouchon - nettoyer* non syntaxiquement lié). La précision supérieure ici laisse, quant à elle, penser que notre méthode gagnerait à prendre en considération des informations syntaxiques. Toutefois, les systèmes automatiques d'annotation syntaxique sont encore trop bruités pour être utilisables sans supervision humaine, et une annotation manuelle est quant à elle à exclure pour un très grand volume de textes. Ici aussi, on doit donc choisir, selon ses objectifs, entre des résultats de qualité élevée et des méthodes d'extraction automatiques ou semi-automatiques.

En résumé, ces tests permettent de tirer les conclusions suivantes : d'une part, pour notre objectif, les modèles statistiques [29], totalement automatiques, ne donnent pas des résultats suffisamment satisfaisants pour pouvoir être utilisés tels quels ou sans supervision humaine *a posteriori* ; d'autre part, l'annotation syntaxique manuelle des paires N-V permet certes d'accéder à de très bons résultats, mais elle est trop coûteuse pour être utilisable sur une très grande quantité de textes, les systèmes automatiques ne pouvant quant à eux pas fournir la même qualité d'annotation. Même en présence d'une éventuelle annotation syntaxique automatique parfaite, cette méthode à base de liens syntaxiques conserve des limites, qui plaident pour l'utilisation de notre technique fondée sur la PLI, malgré son taux de précision actuellement plus faible (pour lequel nous suggérons d'ailleurs par la suite des possibilités d'amélioration). Parmi celles-ci, on peut citer l'absence de verbalisation de la théorie et d'explications de ce qui caractérise un couple qualia par rapport à un couple non qualia, ou encore le fait que les liens qualia ne se limitent pas aux seuls liens sujet, objet et modifieur. Par conséquent, notre méthode d'apprentissage symbolique est un très bon compromis, combinant de bons résultats et une intervention humaine limitée à l'étiquetage sémantique et au choix des exemples positifs et négatifs.

4.3.4 Évaluation linguistique

Cette dernière section est dédiée à une discussion à caractère linguistique des différents résultats que nous avons obtenus, tant lors de la phase d'extraction des paires qualia à l'aide des 9 règles apprises que lors de celle de l'apprentissage de ces dernières. Nous examinons, dans un premier temps, les raisons de la non détection de certaines paires qualia et de la reconnaissance de paires incorrectes par les clauses induites. Nous étudions ensuite les règles générales inférées par notre système de

29. Au moins les modèles simples que nous avons utilisés.

PLI et les patrons qu'elles révèlent pour l'expression du concept de rôle qualia, et comparons ceux-ci avec des observations effectuées manuellement sur le même corpus.

Paires N-V correctement ou incorrectement détectées

Si nous considérons les performances de notre système d'extraction, nous pouvons conclure que ses résultats sont globalement très prometteurs : d'une part, il détecte sur l'ensemble de test la plupart des couples qualia ; les 4 couples non détectés proviennent en fait de constructions très rares dans notre sous-corpus test, qui ne peuvent donc être prises en considération par la méthode d'apprentissage, comme par exemple *prise-relier* dans *la citerne est reliée à l'appareil par des prises*, où un syntagme prépositionnel (SP) *à l'appareil* est inséré entre le V et le SP introduit par la préposition *par* contenant le N. D'autre part, seules 8 paires parmi 36 couples non qualia incorrectement reconnus ne sont pas en relation syntaxique ; l'algorithme de PLI peut donc de manière assez fiable distinguer les paires syntaxiquement liées ou non.

En comparant ces résultats à ceux obtenus par les méthodes statistiques, une conclusion générale assez évidente s'impose : le principal problème pour les méthodes statistiques est le silence, alors que pour notre système, c'est la précision. Cependant, on peut également distinguer deux types d'erreurs parmi celles commises par la méthode de PLI. Le premier est dû à des constructions qui sont effectivement ambiguës, dans lesquelles le N et le V peuvent être ou non syntaxiquement reliés, comme *enlever-prises* dans *enlever les shunts sur les prises*. De tels couples ne peuvent être disambiguïsés par des indices contextuels superficiels tels que les étiquettes, et montrent donc une limitation de l'apprentissage à partir des seules informations catégorielles et sémantiques. Ils sont cependant rares dans notre sous-corpus (8 paires). Au contraire, les autres erreurs semblent davantage liées à des choix faits lors de la mise au point de la méthode d'apprentissage, qui pourraient être facilement modifiés. Ainsi, considérer le nombre de noms entre le V et le N permettrait d'éviter de reconnaître de nombreuses paires telles que *poser-capot* dans *poser les obturateurs capots* ou *assurer-voyant* dans *s'assurer de l'allumage du voyant*.

Après cette discussion rapide des raisons des quelques problèmes de détection de couples qualia, nous nous intéressons maintenant aux règles apprises, outils permettant cette extraction.

Validité linguistique des règles apprises

Pour un linguiste, le problème n'est pas uniquement de détecter de bonnes occurrences de paires en relation qualia, mais également d'identifier dans des corpus les structures linguistiques exprimant ces relations. La question que l'on se pose alors est : qu'est-ce que les clauses inférées nous apprennent sur les structures linguistiques susceptibles de porter un lien qualia entre un nom et un verbe? Des travaux précédents sur d'autres types de liens sémantiques [Mor99] nous ont enseigné que de telles relations pouvaient être portées par une grande variété de structures, et

que ces ensembles de structures pouvaient varier d'un corpus à un autre. Ces recherches se focalisent cependant sur des liens constituant la base d'ontologies telles que l'hyperonymie ou la méronymie. Notre but est similaire, mais avec la difficulté supplémentaire que les relations que nous étudions (liens télique, agentif...) n'ont jamais été étudiées extensivement en corpus et sont plus difficilement identifiables que certaines relations sémantiques conventionnelles.

Nous sommes donc face à l'ensemble de 9 règles apprises par notre système (*cf.* section 4.3.2) que nous cherchons à interpréter de manière linguistique. Au degré de généralisation obtenu dans cette expérience, on constate que peu de traits linguistiques « classiques » sont retenus par les règles. Les clauses semblent fournir des indications très générales mais peu d'informations sur les types de verbes (deux clauses mentionnent par exemple l'étiquette sémantique « verbe d'action »), de noms (nom commun singulier est mentionné) ou prépositions qui apparaissent dans les structures trouvées. Ces clauses contiennent toutefois d'autres informations liées à plusieurs aspects de descriptions linguistiques, comme :

- la *proximité* : c'est un critère majeur. La plupart des clauses indiquent que le N et le V doivent être soit contigus (clause 2), soit séparés par au plus deux éléments (clauses 3, 4, 6, 7) et qu'aucun verbe ne doit apparaître entre le N et le V (clauses 1, 3, 8, 9) ;
- la *position* du N ou du V dans la phrase : les clauses 4, 5, 7 et 8 indiquent qu'un des deux éléments doit se trouver en début de phrase ou immédiatement après un signe de ponctuation, tandis que les positions relatives du N et du V (precede/2) sont signalées dans les clauses 1, 3, 5, 7 et 9 ;
- la *ponctuation* : les signes de ponctuation, plus particulièrement les « : », sont mentionnés en clauses 5 et 7 ;
- la *catégorisation morpho-syntaxique* : la première clause détecte une structure importante correspondant à des verbes d'action à l'infinitif.

Ces traits marquent donc des schémas linguistiques très spécifiques à notre corpus, texte technique contenant des instructions. Dans ce texte apparaissent en effet de nombreux exemples de verbes à l'infinitif, en début de phrases, suivis d'un syntagme nominal (*débrancher la prise, déposer les obturateurs*).

Pour mieux évaluer nos résultats, nous les avons comparés à des observations manuelles réalisées par Édith Galy[30] sur le même corpus [dG00]. Galy a ainsi listé un ensemble de structures verbales canoniques porteuses du lien qualia télique :

- verbe infinitif + det + nom (*visser le bouchon*)
- verbe + det + nom (*ferment le circuit*)
- nom + participe_passé (*bouchon maintenu*)
- nom + être + participe_passé (*circuits sont raccordés*)
- nom + verbe (*un bouchon obture*)
- être + participe_passé + par + det + nom (*sont obturées par les bouchons*)

D'une part nos résultats se recoupent, les deux travaux montrant l'importance des structures infinitives et relevant des patrons dans lesquels le N et le V sont

30. Mémoire de maîtrise en Sciences du langage, Université de Toulouse-Le Mirail.

proches. En fait, notre méthode propose des généralisations des structures découvertes par Galy. En particulier, l'opposition entre constructions passive et active est fusionnée dans la clause 2 par l'indication de contiguïté (le V peut se trouver avant ou après le N). Certains schémas valides ne sont cependant pas retrouvés, en particulier quand les marqueurs sont des expressions polylexicales telles que « avoir pour but de », « être utilisé pour », « avoir pour fonction de »..., puisque ces expressions ne sont pas repérées[31]. En revanche, nous repérons certains indices qui n'ont pas été détectés par l'analyse manuelle, car ils dénotent des niveaux d'information linguistique généralement négligés par l'observation linguistique, telle que la ponctuation et la position dans la phrase.

Quand on considère d'un point de vue linguistique les résultats du processus d'apprentissage, il apparaît donc que les clauses donnent des indices de surface très généraux sur les structures qui, dans le corpus, favorisent l'expression de liens qualia. Ces indices sont suffisants pour donner accès à certains patrons spécifiques du corpus, ce qui constitue un résultat intéressant. Nous revenons en conclusion de ce chapitre (*cf.* section 4.5) sur la possibilité d'obtenir des schémas plus ou moins généraux en paramétrant le *background knowledge*, mais également sur les types de règles inférées lors de nos diverses expériences d'apprentissage manipulant des étiquetages plus ou moins importants (catégoriel seul, catégoriel et sémantique...).

L'apprentissage mené en prenant en compte toutes les informations sémantiques et catégorielles disponibles sur le contexte environnant du couple et sur le couple lui-même conduit donc à des résultats satisfaisants pour l'extraction de relations nomino-verbales, mais également pour révéler certaines structures porteuses de liens qualia. Toutefois, si nous avons, grâce à la définition d'un opérateur de raffinement adéquat, amené le coût calculatoire dû à l'exploitation de toutes ces informations à des proportions correctes, il convient de rappeler le coût humain encore relativement élevé dû, en particulier, à la mise en place de l'étiquetage sémantique des noms qui nécessite l'examen par un expert de tous les mots du corpus[32]. Nous avons donc cherché à limiter celui-ci en étudiant les résultats qu'il est possible d'obtenir sans prendre en compte, lors de l'apprentissage, les étiquettes sémantiques des noms.

4.4 Apprentissage sans prise en compte de l'étiquetage sémantique des noms

Notre objectif étant d'automatiser l'extraction de couples qualia à partir de corpus, afin de pouvoir disposer de telles ressources pour des besoins applicatifs, il convient donc que notre méthode d'apprentissage de règles permettant d'extraire ces paires nomino-verbales soit la plus portable possible. Nous souhaitons par conséquent tester s'il est possible de ne pas tenir compte de l'étiquetage sémantique des noms tout en obtenant de bons résultats d'apprentissage, cet étiquetage étant le plus

31. Elles ne sont par exemple pas étiquetées en tant que telles.

32. Plusieurs journées de travail (temps de familiarisation avec le corpus et le domaine d'étude inclus) ont été nécessaires à une personne pour réaliser la classification des noms du corpus MATRA-CCR.

coûteux en temps lorsque l'on passe d'un corpus d'un domaine à un autre. Nous nous intéressons ici à une approche, dite hybride, dictée par un souci ce portabilité. Les exemples sont codés de la même manière que dans l'expérience précédente et l'opérateur de raffinement est identique. Nous ôtons simplement du langage d'hypothèses les prédicats correspondant aux étiquettes sémantiques des noms. Ceci signifie que la généralisation des exemples ne peut plus se faire sur les catégories sémantiques des noms apparaissant dans le contexte de couples N-V, ni sur la catégorie sémantique du N. Les informations de nature sémantique sur les verbes, les prépositions et les autres catégories de mots du corpus sont en revanche conservées.

Ce choix s'explique, comme nous venons de le mentionner, par le fait que, contrairement à l'étiquetage catégoriel qui nécessite peu d'intervention humaine et reste relativement portable d'un corpus à l'autre, l'étiquetage sémantique est coûteux et peu adaptable d'un corpus à un autre, notamment à cause de la construction du lexique des noms qui est la catégorie apportant le plus d'ambiguïtés (voir section 4.1.2). Nous voulons donc ici confronter notre méthode d'apprentissage à un corpus qui ne comporte que des informations pouvant être ajoutées de manière quasi automatique et peu coûteuse.

Nous présentons ci-dessous et discutons les résultats de l'apprentissage réalisé avec Aleph dans ces conditions. Le tableau 4.6 précise les résultats de la validation théorique de l'apprentissage.

	Temps (secondes)	Précision (%)	Rappel (%)	Coefficient de Pearson
Moyenne	2146	83.4	89.6	0.722
Écart-type	1624	4.6	2.3	0.043

TAB. 4.6 – *Résultats de la validation croisée*

On constate une légère amélioration des taux de rappel et précision et, par conséquent du coefficient de Pearson. Cette différence peut s'expliquer par le fait que l'expérience précédente manque peut-être d'un nombre suffisant d'exemples positifs en regard de la précision de description disponible pour l'algorithme de PLI. En effet, l'utilisation des informations sémantiques sur les noms implique d'ajouter au langage d'hypothèses 33 nouveaux mots correspondant aux 33 étiquettes sémantiques des noms. La granularité trop fine des hypothèses qui en résulte peut conduire l'algorithme à apprendre « par cœur » les E^+ (on parle alors d'*overfitting*, les règles générées reprenant en partie les exemples sans les généraliser) faute de pouvoir trouver des régularités dans les exemples trop peu nombreux.

Le processus d'apprentissage, dans ces nouvelles conditions, produit 7 clauses généralisées :

(1) est_qualia(A,B) :- precede(B,A), proche_verbe(A,B), infinitif(B), verbe_action(B).
(2) est_qualia(A,B) :- contigu(A,B).
(3) est_qualia(A,B) :- proche_verbe(A,B), suc(B,C), preposition(C), pred(A,D), vide(D).
(4) est_qualia(A,B) :- precede(B,A), proche_mot(A,B), proche_verbe(A,B), suc(B,C), preposition(C).

(5) est_qualia(A,B) :- precede(A,B), proche_mot(A,B), pred(A,C), ponctuation(C).
(6) est_qualia(A,B) :- precede(A,B), proche_mot(A,B), suc(A,C), verbe(C).
(7) est_qualia(A,B) :- suc(B,C), suc(C,D), suc(D,A), conjonction_coordination(C).

La validation empirique réalisée dans les mêmes conditions que pour l'expérience précédente conduit aux résultats résumés dans le tableau 4.7.

	Rappel (%)	Précision (%)	F-mesure	Coefficient de Pearson
Système de PLI	92.0	65.0	0.76	0.694

TAB. 4.7 – *Résultats empiriques de la méthode d'apprentissage de type PLI*

Ces données chiffrées sont légèrement supérieures à celles obtenues en prenant en compte les informations sémantiques des noms (*cf.* section 4.3.2). L'extraction de paires qualia sans considérer cet étiquetage des noms semble conduire à de meilleurs résultats, alors que, comme nous l'avons mentionné en page 60, certaines structures catégorielles identiques porteuses ou non de couples N-V qualia peuvent être différenciées uniquement si on traite l'étiquetage sémantique des noms. Cela peut toutefois là encore s'expliquer par la raison évoquée pour justifier la différence notée lors de l'évaluation théorique, et donc par le fait que l'apprentissage de l'expérience précédente a été de moins bonne qualité (Pearson de 0,69 contre 0,72 ici), ce qui s'est donc traduit par la génération de clauses moins intéressantes qu'elles n'auraient pu l'être. Ce dernier apprentissage ne permet donc pas réellement de remettre en cause l'utilité des informations sémantiques, mais il soulève un problème important d'alimentation en exemples de notre système d'acquisition de couples qualia. Avant de pouvoir réellement conclure sur cette l'(in-)utilité de l'étiquetage sémantique des noms pour notre corpus, il convient de réaliser de nouveaux tests avec plus d'exemples.

Les 7 clauses obtenues en n'utilisant que les informations catégorielles des mots et les informations sémantiques sur les catégories de mots autres que les noms reprennent pour la plupart des schémas linguistiques donnés en section 4.3.2. Une fois de plus, nous notons l'importance du critère de proximité. On voit aussi émerger une clause (7) qui permet de détecter des liens qualia entre un N et un V apparaissant dans une liste de prédicats verbaux liés par une conjonction de coordination. C'est par exemple le cas du couple *contre-écrou desserrer* extrait par cette clause dans la phrase « *Desserrer et déposer le contre-écrou avec rondelle du support.* » et du couple *tuyauterie brancher* dans « *Brancher et serrer les tuyauteries : refoulement pompe, retour by-pass...* ».

Cette dernière expérience donne donc, comme la précédente, de bons résultats, à la fois en termes de production de règles linguistiquement pertinentes et en termes de construction de lexiques de couples qualia. Elle présente cependant en plus l'avantage d'être relativement portable d'un corpus à l'autre puisque les informations apportées au corpus peuvent l'être de manière quasi automatique.

4.5 Bilan et discussions

Pour clore ce chapitre, nous présentons un bilan de nos contributions, d'une part par rapport à notre tâche d'extraction de liens nomino-verbaux qualia, d'autre part sur les trois volets d'apprentissage, linguistique et applicatif de notre travail. Ce bilan est également l'occasion pour nous de dégager, sur certains de ces points, – et, en particulier le dernier que nous n'avons pour l'instant que très peu abordé – des perspectives de travail, et de discuter la possibilité d'obtenir des règles contenant davantage d'éléments linguistiques « traditionnels » (type de prépositions...) en limitant la généralisation des clauses produites par l'algorithme de PLI.

Si nous laissons momentanément de côté le caractère linguistiquement motivé des règles d'extraction inférées, qui était aussi un des buts que nous nous étions fixés et dont nous discutons plus loin, notre objectif initial était double : d'une part, mettre au point une méthode d'apprentissage automatique fiable et robuste de règles permettant l'acquisition de paires N-V qualia, d'autre part limiter le coût de cette méthode afin de faciliter son portage d'un corpus à un autre. Les résultats que nous avons présentés dans les sections précédentes permettent de conclure que notre méthode à base de PLI répond très correctement au premier critère. Nous avons, en particulier, montré qu'elle constitue un bon compromis, combinant de bons résultats d'extraction de couples N-V et une intervention humaine limitée, lorsqu'on la compare à des méthodes statistiques ou syntaxiques (*cf.* section 4.3.3). Si, lors de notre évaluation empirique, le taux de rappel est supérieur à 90%, il est également possible de faire, si besoin, croître la précision (actuellement de l'ordre de 65%) [33], d'une part en contrôlant le nombre de détections d'occurrences d'une même paire par les règles d'extraction apprises (facteur x, *cf.* section 4.3.2), d'autre part en travaillant sur le paramétrage de l'algorithme de PLI (*cf.* l'exemple cité en section 4.3.4, consistant à prendre en compte le nombre de noms entre le N et le V dans une phrase).

Pour ce qui est du coût de la méthode, qui doit bien évidemment être minimisé si l'on veut effectivement la porter sur d'autres corpus [34], deux facteurs sont à considérer : l'étiquetage sémantique du corpus et la constitution des exemples d'apprentissage. Pour ce qui concerne ce premier point, et même si des tests supplémentaires sont nécessaires, les résultats obtenus par la méthode hybride que nous avons proposée (*cf.* section 4.4) montrent qu'il semble raisonnable, pour une tâche de constitution automatique de couples N-V qualia, de ne pas prendre en compte l'étiquetage sémantique coûteux des noms. Cette méthode hybride a également mis encore davantage en lumière l'impact du second facteur de coût de notre technique d'apprentissage supervisé, c'est-à-dire les exemples à fournir en entrée de l'algorithme de PLI. Nous avons commencé à étudier les potentialités de combinaisons de méthodes d'apprentissage pour tenter de limiter le nombre d'exemples nécessaires

33. Dans le cadre, par exemple, d'une application où la qualité des paires obtenues ne peut être validée manuellement mais doit être attestée.

34. Nous avons ainsi recommencé notre apprentissage sur un corpus moins technique, qui contient des textes de la Fortisbank de Bruxelles décrivant la mise en place de l'Euro et ses conséquences pour des particuliers.

[Cat02]. Nous nous sommes plus particulièrement intéressée jusqu'à présent à la technique de *co-training* proposée par Blum et Mitchell [BM98] ; l'exploration de ces combinaisons constitue l'une de nos perspectives de travail.

Nous avons déjà développé en section 4.2 nos contributions dans le cadre de l'apprentissage et nous contentons donc de les rappeler ici très brièvement. Elles portent essentiellement sur la définition d'un opérateur de raffinement bien adapté aux connaissances hiérarchisées que nous manipulons, qui permet de parcourir de manière efficace le treillis d'hypothèses organisé selon un ordre de généralité que nous avons précisé (la θ_{NV}-subsomption), et produit des règles linguistiquement motivées et bien formées en évitant de parcourir, grâce à l'utilisation de propriétés privées et d'une fonction de score que nous avons définie, des branches inutiles du treillis. Ce travail permet à ALEPH de traiter en des temps corrects une somme très importante de connaissances (étiquettes catégorielles et sémantiques des mots présents dans les phrases d'où sont issus les couples N-V servant à constituer les E^+ et E^-).

Les contributions sur le plan linguistique de nos travaux concernent la mise en évidence de patrons, propres à chaque corpus, qui favorisent l'expression de liens qualia, ce qui constitue en soi un apport à la théorie de LG qui ne connaît actuellement pas les schémas caractéristiques des divers rôles. Contrairement aux travaux déjà cités de Pustejovsky *et al.* [PAB93] qui cherchent à identifier des mots liés par ces rôles au sein de relations syntaxiques pré-spécifiées, nous n'avons aucun *a priori* sur les structures révélatrices des liens qualia ; la méthode d'apprentissage explicative que nous avons développée met au jour des structures très spécifiques du corpus manipulé. Comme nous l'avons mentionné en section 4.3.4, au niveau de généralité des clauses atteint par la version de l'apprentissage décrite ici, peu d'indices linguistiques « traditionnels » subsistent dans ces patrons ; nos travaux suggèrent l'importance de niveaux d'informations linguistiques souvent ignorés par l'observation manuelle habituelle tels que la position, la proximité..., ainsi que l'intérêt de détecter des patrons spécifiques à chaque corpus.

Notre objectif a, jusqu'ici, été de définir une technique d'apprentissage permettant d'acquérir des règles les plus pertinentes possibles pour extraire des couples N-V qualia, ce qui nous a conduite à opposer, lors de l'apprentissage, le concept de paire qualia à celui de paire non-qualia sans distinction des rôles, et à maximiser la généralisation contrôlée des clauses apprises. Les recherches que nous avons effectuées peuvent également être orientées vers un but linguistique de verbalisation la plus lisible possible de la théorie du LG, de recensement d'un plus grand nombre de structures caractérisant les rôles qualia selon les corpus, et de distinction des patrons propres à chaque rôle. Nous examinons successivement ces trois aspects en débutant par le dernier.

Si cela est souhaité, notre méthode d'apprentissage peut être utilisée pour produire des clauses caractérisant chacun des rôles qualia ; il suffit pour ce faire de fournir en entrée de l'algorithme d'apprentissage suffisamment d'exemples de chaque type. Pour augmenter la couverture des schémas porteurs des rôles, on peut, par exemple, chercher à intégrer, lors de l'apprentissage, la reconnaissance et la prise en

compte de certaines expressions polylexicales telles que celles relevées manuellement par É. Galy. Enfin, en ce qui concerne la verbalisation de la théorie et des divers rôles qualia, il est possible de limiter la généralisation (et donc le degré d'abstraction) des règles produites par l'algorithme de PLI, en particulier en influant sur le contenu du *background knowledge*, afin d'obtenir des clauses laissant, par exemple, apparaître davantage d'éléments linguistiques « habituels » (type de prépositions...). En effet, nos premières expériences d'apprentissage, recensées dans l'introduction de ce chapitre page 60 et décrites dans [CSBF01], laissaient apparaître de tels éléments dans les clauses obtenues avec un algorithme moins optimisé[35]. Les divers apprentissages réalisés – certes, dans des conditions parfois différentes – en prenant en compte des niveaux d'étiquetage variés invitent à une réflexion sur le lien existant entre les étiquettes considérées lors de l'inférence des règles et le type de clauses et de schémas linguistiques révélés.

Le premier apprentissage, réalisé sur le corpus MATRA-CCR annoté uniquement par des étiquettes catégorielles, a ainsi déjà mis en évidence des clauses généralisées dénotant l'importance des critères de proximité, de position et le rôle de la ponctuation. Il a aussi permis de montrer la pertinence de certaines constructions syntaxiques. L'une d'entre elles caractérise la tournure passive : N et V sont en relation si le V est précédé de l'auxiliaire *être* et suivi de la préposition *par*. Deux clauses apprises spécifient, de manière *a priori* plus surprenante, que le N et le V qui suivent une conjonction de subordination sont pertinents[36], ce qui généralise le fait que beaucoup de verbes, dans le corpus MATRA-CCR, requièrent des complétives indiquant une action typique, comme « s'assurer que » ou « vérifier que » (« s'assurer que l'*alimentation* est *coupée* » ; « vérifier que le *feu* anti-collision *clignote* »).

Le second apprentissage, basé sur l'exploitation des étiquettes catégorielles et sémantiques de mots proches des N et V des exemples et de l'étiquette catégorielle du V, conduit certes à un certain nombre des clauses déjà obtenues à l'aide de l'étiquetage catégoriel, mais induit également des règles spécifiant des propriétés sémantiques intéressantes sur les mots du contexte du couple N-V. Ainsi, une clause précise que les verbes modaux comme *permettre*, *devoir* ou *pouvoir* sont de bons indicateurs de couples pertinents : « le *tableau* doit être *éclairé* », « l'*adhésion* peut être *atteinte* »... ; une autre indique que le type sémantique de la préposition peut aider à identifier des couples qualia, spécialement si la préposition indique la manière ou le but : « *fixer* avec leurs *vis* sans serrer ».

Les expériences prenant en compte l'intégralité des étiquetages catégoriel et sémantique ou laissant de côté les étiquettes sémantiques des noms décrites en sections 4.3 et 4.4, en plus de l'inférence de règles marquant aussi l'influence des éléments de ponctuation, position et proximité, font apparaître des clauses où l'importance du type sémantique du verbe est visible. Une version de l'algorithme d'apprentissage conduisant à moins de généralisations [CSBF01] produit aussi, pour la première d'entre elles, des clauses dénotant une structure passive de type télique dans laquelle l'étiquette sémantique du N instrument (événement) est précisée ; une autre

35. Et, par conséquent, produisant des règles moins pertinentes en termes d'extraction de couples qualia.

36. On retrouve d'ailleurs ce cas dans la clause 9 de la page 74.

clause met en évidence des schémas désignant une action typique à produire sur un objet exprimés par la suite « *en* + participe présent + prép » ; quelques clauses caractérisent avec plus de précision par un filtrage sémantique certaines des relations mises au jour lors des deux premières expériences, par exemple entre la préposition *sous* et un objet de type *état* (*sous tension*...).

Au chapitre 2, nous avons indiqué un certain nombre d'arguments en faveur des liens inter-catégoriels N-V de type qualia pour caractériser des variantes sémantiques de noms et les désambiguïser. Nous avons également mentionné que, par conséquent, ces paires N-V qualia étaient potentiellement pertinentes pour accroître les performances de systèmes de recherche d'information (SRI). Nos travaux visent donc aussi à contribuer à des applications de type RI par l'exploitation d'un lien nomino-verbal original et très peu usité pour l'extension de requêtes et la désambiguïsation des questions. Nous débutons seulement l'étude de l'influence exacte, pour un SRI, de la prise en compte des paires N-V qualia acquises en corpus à l'aide de notre méthode d'apprentissage, et ne pouvons donc présenter actuellement des données chiffrées ou des contributions attestées sur ce point. Nous nous limitons donc, pour terminer cette section, à une brève présentation de travaux préliminaires que nous avons déjà réalisés sur le sujet et de perspectives à court terme que nous envisageons.

Lors de sa thèse réalisée sous mon encadrement, Cécile Fabre [Fab96] a montré l'intérêt des prédicats verbaux qualia pour interpréter les séquences binominales ne contenant pas de déverbal. Ainsi, ce sont des rôles qualia qui permettent de comprendre une structure *N Prép (Dét) N* telle *couteau à pain* (prédicat *couper*, rôle télique) ou *message du compilateur* (prédicat *émettre*, rôle agentif). En supposant résolu le problème de l'association de prédicats qualia aux noms, les étapes du calcul de la sémantique d'une séquence binominale sont alors les suivantes : détermination du ou des prédicats associés aux constituants en se focalisant essentiellement sur les prédicats associés au nom tête, filtrage des schémas prédicatifs effectivement possibles pour la séquence en se basant sur des contraintes de typage sémantique associées aux arguments des prédicats, et, pour les structures complexes en français, sur le rôle sémantique de la préposition et du déterminant. L'insertion[37] de ces mécanismes dans le cadre d'un SRI des services télématiques du CNET[38] [FS99] nous a permis de montrer que[39] le contexte de la séquence binominale pouvait être utilisé pour désambiguïser les noms à condition que des liens syntagmatiques N-V qualia soient développés dans le thésaurus du système et exploités par une fonction de désambiguïsation. Elle a également permis d'avoir accès à des liens de paraphrase sémantique entre requêtes et textes : nous avons montré que l'appariement devait favoriser les textes contenant le concept complexe exprimé par la requête via des prédicats qualia plutôt que ceux contenant un concept associé à un seul des constituants des séquences.

Nous avons également cherché à évaluer l'intérêt de l'utilisation de N-V qualia dans le cadre de requêtes effectuées par des documentalistes du département

37. Partielle pour cause de contraintes inhérentes au système à notre disposition pour ce test.
38. Maintenant France Télécom R & D.
39. Les requêtes lors de ce test étaient toutes de la forme *N Prép (Dét) N*.

documentation de la banque Fortisbank à Bruxelles. L'objectif de l'enquête réalisée par Laurence Vandenbroucke [40] [Van00] était de tester si ces couples pouvaient effectivement leur servir à désambiguïser leurs requêtes et préciser les documents pertinents pour eux. Via une interface d'une part et un questionnaire papier d'autre part, chaque documentaliste devait pointer, parmi des verbes entretenant un lien qualia ou non avec des noms qu'il souhaitait utiliser dans une requête, ceux qui lui permettraient d'obtenir au plus vite les documents effectivement recherchés. Ainsi, plutôt que poser des questions très vagues telles que « *contrat* », l'utilisateur pouvait se servir de verbes « *établir* » ou « *résilier* » pour préciser son but. Cette enquête, même si elle n'a pu être menée qu'auprès de quelques documentalistes, a montré que les liens qualia repéraient les occurrences des noms qui les intéressaient effectivement, aucune des paires N-V non qualia proposées lors de l'enquête ne leur ayant semblé intéressantes.

Nous débutons actuellement l'insertion de mécanismes de prise en compte de paires N-V qualia dans un SRI pour évaluer de manière plus systématique leur apport, sur les données de la seconde campagne Amaryllis. Notre objectif est plus particulièrement d'exploiter ces relations nomino-verbales sur les champs sujet et concept des requêtes, en les étendant à l'aide de couples N-V acquis en répétant notre apprentissage sur le corpus du quotidien *Le Monde* de cette campagne. Nous présentons cependant en conclusion générale d'autres modes d'évaluation des apports de ressources linguistiques aux SRI que nous voulons mettre en place, pour ne pas nous restreindre à une évaluation dans des cadres présupposant une liste de réponses à découvrir et ne prenant pas en compte la pertinence des documents retournés pour un utilisateur effectif.

L'insertion des relations qualia dans un SRI soulève toutefois de nouvelles questions. Par exemple, il convient de se pencher sur les transformations morpho-syntaxiques à faire subir à une requête complexe (*jaugeur de carburant*) lorsqu'on l'étend à l'aide d'un prédicat (*mesurer* du *carburant*). Les travaux de Chr. Jacquemin *et al.* [Jac01, JT99, FJ00] peuvent donner quelques pistes intéressantes à ce sujet. D'autre part, il faut également se poser la question du choix des prédicats verbaux qualia à utiliser pour étendre une requête : pour un nom donné, sur un corpus donné, les règles que nous inférons par PLI extraient un nombre potentiellement élevé de verbes, et ce, pour chaque rôle qualia. Si tous ces verbes correspondent effectivement à un des rôles qualia de ce nom dans le corpus, tous les liens retenus n'ont peut-être pas la même pertinence en RI. Ainsi, sur un corpus, on peut très bien avoir acquis à la fois *enseigner* comme prédicat télique de *livre*, mais aussi *caler* si ce livre y sert à caler une table. Ce second verbe peut sembler *a priori* moins intéressant à utiliser de manière systématique pour étendre des questions. Une notion de typicalité des prédicats verbaux qualia semble donc à étudier. Cependant, il faut également se rendre compte que certains éléments très spécifiques, évoquant une facette particulière d'un nom, peuvent aussi être particulièrement pertinents pour discriminer des documents.

40. Ce travail a été effectué dans le cadre de son stage de DES information et documentation (Université libre de Bruxelles).

Nous venons de lister quelques-unes des pistes de travail que nous souhaitons explorer. Dans le chapitre conclusif suivant, qui est pour nous l'occasion de faire aussi un point plus général sur notre travail et son bilan, nous résumons nos perspectives à court et plus long termes.

Chapitre 5

Conclusions et perspectives

Dans ce document synthétisant nos travaux sur l'acquisition de relations lexicales sémantiques en corpus, nous avons pris le parti d'effectuer, à l'issue de chaque chapitre, un bilan du travail réalisé (*cf.* sections 3.2.2 et 4.5) ; nous avons aussi précisé en introduction la position et les particularités de nos recherches par rapport aux autres études réalisées dans le domaine. Nous ne reprenons donc ici, en guise de conclusion, que brièvement certains des points déjà énoncés, et nous citons ensuite des perspectives que nous souhaitons explorer.

5.1 Bilan

La spécificité de nos recherches par rapport à d'autres études menées sur l'acquisition d'informations sémantiques en corpus concerne notre positionnement original par rapport à ce problème, que nous considérons sous un triple aspect linguistique, applicatif et d'apprentissage. Ceci se décline en particulier par le fait que nous utilisons des cadres linguistiques formels pour définir des informations pertinentes pour une application visée, pour guider la mise au point de nos méthodes d'apprentissage et pour contrôler la validité de ce que nous apprenons. De plus, nous cherchons à travailler effectivement sur les trois facettes du problème.

Par rapport à notre tâche d'enrichissement de la description lexicale des noms dans une optique de désambiguïsation et de prise en compte des variantes sémantiques, nos contributions portent sur deux points : d'une part, nous étudions un type de lien transcatégoriel très peu sollicité jusqu'à présent, que nous acquérons par une méthode d'apprentissage symbolique explicative ; d'autre part, nous ne considérons pas uniquement des liens intracatégoriels nominaux « traditionnels », mais cherchons à apprendre de manière la plus automatique possible des liens sémiques qui ont des potentialités intéressantes en termes d'exploration de variations sémantiques (*cf.* section 5.2). En ce qui concerne l'apprentissage des liens N-V qualia, nous devons chercher à déterminer des techniques permettant de réduire encore le coût de l'apprentissage supervisé, et réeffectuer l'intégralité des apprentissages sur plusieurs corpus (tâche en cours sur deux corpus) pour pouvoir discuter plus en avant la

diversité des clauses apprises selon le type du corpus. De même, les phases deux et trois de la méthodologie d'acquisition de liens sémiques que nous avons décrite au chapitre 3 doivent encore être fortement travaillées pour aboutir à une mécanisation effective de la production de taxèmes structurés par des sèmes. Sur le plan linguistique, outre le fait que nous montrons la faisabilité[1] du développement à grande échelle de lexiques respectant certains des principes des théories linguistiques retenues, nous apportons une pierre à la réflexion linguistique sur les rôles définis dans la structure des qualia du LG, grâce à la production de règles permettant de verbaliser la théorie et de déterminer, selon les corpus étudiés, les structures susceptibles de les caractériser. En ce qui concerne l'apprentissage, nous avons enrichi des techniques existantes pour les adapter à notre problème, en proposant une méthode de densification de matrices creuses et de remise en cause partielle d'agglomérations effectuées par un algorithme de classification hiérarchique, et en définissant un ordre de généralité du treillis des hypothèses exploré par l'algorithme de PLI que nous manipulons et un opérateur de raffinement pertinents pour gérer des connaissances hiérarchisées. Sur le plan applicatif, notre apport réel est actuellement non quantifiable puisque nous débutons uniquement la prise en compte de liens N-V qualia au sein d'un SRI, et que nos travaux sur les liens N-N ne sont pas suffisamment aboutis pour pouvoir tester l'insertion de liens sémiques. Cependant, nous nous attachons à exploiter en RI des relations intra- et intercatégorielles qui ont jusqu'à présent été très peu, voire pas, utilisées dans ce cadre.

Ce dernier point révèle un des futurs axes de notre travail, que nous présentons maintenant.

5.2 Perspectives

Apprentissage d'informations sémantiques en corpus

Parmi les perspectives que nous envisageons de développer, l'une d'elles concerne la poursuite de nos travaux sur l'acquisition de connaissances lexicales sémantiques en corpus, et l'exploration des pistes qu'ils ont laissé entrevoir. Nous ne listons pas ici l'intégralité des points à étudier, mais nous focalisons sur trois d'entre eux.

Nous avons, dans ce document, mentionné la nécessité de chercher à augmenter la portabilité des méthodes d'apprentissage des éléments linguistiques en corpus, en limitant l'investissement nécessaire pour passer de leur application d'un corpus à un autre. Dans cet objectif, il nous paraît intéressant d'étudier l'apport des combinaisons de méthodes d'apprentissage, tant statistiques que symboliques. Nous avons déjà effleuré ce sujet en testant l'intérêt du *co-training* pour limiter le nombre d'exemples étiquetés à fournir en entrée d'un algorithme d'apprentissage supervisé. Nous allons explorer diverses combinaisons et évaluer les bénéfices, en termes de coût, de qualité d'apprentissage..., qu'elles permettent d'envisager selon le problème de TAL traité (acquisition d'informations en corpus, mais aussi étiquetage sémantique...).

1. Au moins dans une certaine proportion pour la SD.

Un second sujet porte sur l'utilisation combinée de plusieurs ressources linguistiques pour acquérir des relations sémantiques ou, de manière plus générale, des lexiques sémantiques. Jusqu'ici, nous nous sommes intéressée à l'extraction de relations sémantiques à partir des seuls corpus textuels. Si nous maintenons bien sûr que seuls les corpus du domaine étudié peuvent permettre d'acquérir les façons effectives dont les mots sont utilisés dans ce domaine, il peut être intéressant de se pencher sur ce qu'une combinaison d'informations extraites à partir de corpus et de celles obtenues à l'aide d'autres ressources, comme des dictionnaires ou des thésaurus par exemple, peut apporter à la description des mots. Ainsi, dans une optique de constitution de représentations lexicales basées sur le Lexique génératif de Pustejovsky auxquelles nous nous intéressons, l'exploitation par apprentissage de dictionnaires peut donner accès à des structures des qualia par défaut, l'acquisition en corpus servant alors à compléter la liste des prédicats pouvant jouer les divers rôles qualia d'un nom donné. Nous désirons tester si la combinaison des diverses sources de connaissances permet d'avoir accès à tous les aspects sémantiques potentiellement intéressants, tout en réfléchissant à ce que chacune apporte par rapport aux autres.

Le troisième point concerne l'étude de la variation sémantique proprement dite, en exploitant des liens du type de ceux que nous acquérons. Plus précisément, nous voulons explorer la variation de sens induite par le remplacement d'un mot par un quasi synonyme ou celle obtenue lors de modifications dues à un lien qualia. Nous souhaitons en particulier tester si des représentations de la granularité de celles présentes dans les lexiques basés sur la SD peuvent être utilisées en ce sens. De la même façon et avec la même rigueur que [FJ00] définit des contraintes précises pour qu'une forme puisse être considérée comme une variante d'un terme basée sur une relation nomino-verbale contenant un lien morphologique, nous voulons évaluer dans quelle mesure, potentiellement modeste, il serait possible de contrôler la modification d'un terme sur le plan sémantique, tout en référant toujours le « même » concept. Parmi les nombreuses pistes à envisager, nous pouvons, par exemple, tester s'il est possible de typer les sèmes devant être conservés entre le terme étudié et sa variante sémantique pour ne pas « trop » s'éloigner du concept initial. Cette étude sur la variation de sens a des applications immédiates en RI (extension de requêtes), mais aussi en génération de textes ou en aide à la traduction.

La PLI pour répondre à des besoins du TAL

Au cours de nos travaux, nous avons pu (très partiellement) explorer l'intérêt de la PLI pour le TAL. Depuis environ cinq ans, cette technique d'apprentissage commence à être utilisée par la communauté *TAL et apprentissage*, essentiellement pour produire des étiqueteurs catégoriels et des analyseurs morphologiques et syntaxiques. Parmi les quelques travaux, plus proches de nous, exploitant la PLI pour des besoins plus sémantiques, on peut citer [Néd99] qui apprend, à partir de corpus spécialisés annotés syntaxiquement, des cadres de sous-catégorisation et des classes conceptuelles, et [Moo99] qui acquiert des lexiques (sous la forme mot - représentation logique) et des analyseurs sémantiques permettant de produire des représentations de phrases sous la forme d'expressions logiques.

La PLI est une technique très prometteuse pour répondre à certains besoins du TAL. Outre sa pertinence pour extraire des paires N-V qualia à partir de corpus, une autre de ses caractéristiques essentielles nous a particulièrement intéressée lors de nos travaux : la production de règles explicatives. En effet, un des attraits majeurs de la PLI est le recours à un langage d'hypothèses expressif qui en fait un outil facilitant l'échange entre la communauté *TAL et apprentissage* d'une part et la communauté des linguistes d'autre part. Elle permet d'envisager une façon de voir l'acquisition automatique d'informations sur corpus qui soit parlante pour des linguistes. Nous avons jusqu'à présent exploité cette méthode pour apprendre des relations lexicales peu étudiées et peu évidentes à identifier. Nous souhaitons l'utiliser pour d'autres acquisitions, dont, par exemple, celles de relations lexicales sémantiques plus « classiques ».

Insertion de ressources linguistiques dans un SRI

Une troisième perspective concerne l'insertion effective de ressources linguistiques dans des SRI. Nous avons déjà eu l'occasion de mentionner les résultats parfois mitigés auxquels de tels ajouts de connaissances conduisaient (*cf.* section 2.1). Nous partageons donc les avis de Spärck Jones [SJ99] ou de Jacquemin [Jac00] qui affirment qu'il convient de mener une réflexion approfondie sur la façon d'utiliser les ressources issues du TAL et de la linguistique au sein de SRI, en particulier sur les tâches d'accès au contenu que ces informations sont effectivement à même de faciliter. Dans ce vaste débat, nous souhaitons, entre autres, aborder la question du choix, parmi les ressources d'un type donné disponibles, de celles qui sont effectivement à utiliser. Nous avons déjà soulevé ce problème en constatant que, parmi les prédicats remplissant le même rôle qualia pour un nom donné, certains semblaient plus susceptibles que d'autres d'être choisis pour étendre des requêtes. Le problème se retrouve au niveau de liens paradigmatiques où l'extension d'une question à l'aide de mots synonymes peut conduire à des documents non pertinents à cause de la (légère) différence de sens entre les deux éléments. D'où l'intérêt de chercher à déterminer à un niveau de granularité fin les modifications de sens induites. Ceci nous amène naturellement à évoquer un autre aspect fondamental que nous voulons explorer : celui de l'évaluation de l'apport de ressources linguistiques à un SRI. Contrairement aux campagnes d'évaluation telles que TREC ou Amaryllis qui supposent la connaissance *a priori* des documents répondant « effectivement » à une requête, sans tenir compte de l'intérêt qu'un usager réel leur attribuerait et attribuerait à un autre document de la base interrogée, nous voulons établir une méthodologie d'évaluation des apports linguistiques pour des moteurs de recherche fondée sur la satisfaction d'un utilisateur. Nous envisageons de développer une méthode générique, adaptable à des SRI portant sur un domaine spécifique mais également à des moteurs non spécialisés du Web, qui soit capable de gérer des extensions linguistiques de types divers (sémantiques, mais également morphologiques, syntaxiques...) [2]. Pour ce faire, nous proposons d'inférer la satisfaction de

2. Nous avons déjà eu l'occasion d'établir des contacts en ce sens avec l'Irit, l'Erss, le Clips-Imag, l'Inria-Lorraine, et des établissements travaillant sur l'ergonomie des interfaces et leur efficacité en

l'utilisateur à l'aide d'une technique d'apprentissage automatique s'appuyant sur les comportements de « cobayes » donnant explicitement leur avis sur la pertinence de documents proposés lors d'un scénario de recherche. Nous voulons donc apprendre la correspondance entre la satisfaction d'un utilisateur et différents indicateurs (métriques) de son comportement d'interrogation et de navigation et de documents qu'il a annotés comme intéressants, dans deux cas : requêtes automatiquement étendues par les données linguistiques et requêtes non étendues. Notre approche a pour but d'établir des modèles « utilisateurs satisfaits », qui servent ensuite à mesurer automatiquement la satisfaction d'un nouvel utilisateur de manière non intrusive.

Accès au contenu de documents multimédias - couplage texte-image

La dernière perspective que nous mentionnons concerne l'accès au contenu de documents multimédias. Nous souhaitons explorer les potentialités du couplage texte-image[3], en considérant des documents qui contiennent à la fois des images et du texte en forte corrélation (les images ne doivent pas être simplement décoratives) : bases bibliographiques scientifiques, presse, documentation technique... Notre but est d'étudier les apports mutuels entre les deux médias pour une meilleure description de ces documents, et en particulier des images. Ceci suppose tout d'abord de trouver dans le texte les parties qui ont trait aux images, et d'associer aux images une description issue des éléments qui auront pu être extraits du texte. Ce premier lien établi, on peut alors chercher à relier des documents entre eux : documents contenant les mêmes images ou des images proches, documents abordant un même thème. Ces rapprochements visent à accumuler de l'information pour enrichir la description textuelle des images, mais aussi pour aider à désambiguïser le texte et les descriptions des images. Côté visuel, la recherche d'images proches peut être difficile, car la probabilité d'obtenir des images visuellement proches mais sémantiquement très différentes reste grande. D'un autre côté, le fait de trouver dans deux documents la même image – ce qui est assez facile d'un point de vue image – permet de faire des associations entre textes, que la seule utilisation du média texte aurait pu avoir des difficultés à réaliser. Par ailleurs, on peut chercher à aller plus loin que la simple description juxtaposée du texte et de l'image en cherchant une description qui mèle les deux médias. Une piste pour ce faire serait de détecter les associations possibles entre descripteurs de textes et descripteurs d'images. Nous souhaitons tester les capacités de la PLI pour effectuer une telle recherche, afin de comprendre les associations produites. Nous espérons ainsi aller plus loin que les systèmes actuels d'association et de propagation de mots-clés à des images qui sont basés sur des techniques statistiques de partitionnement.

fonction de l'expertise de l'utilisateur, tels que le Centre de recherches en psychologie, cognition et communication de l'Université de Rennes 2.

3. Nous travaillons à l'Irisa au sein de l'équipe TexMex (techniques d'exploitation de documents multimédias: exploration, indexation et recherche dans de très grandes bases), qui regroupe des spécialistes de différents médias.

Bibliographie

[AB96] Houssem Assadi and Didier Bourigault. Acquisition et modélisation de connaissances à partir de textes : outils informatiques et éléments méthodologiques. In *10^e Congrès reconnaissance des formes et intelligence artificielle, RFIA-96*, Rennes, France, 1996.

[ABR95] Susan Armstrong, Pierrette Bouillon, and Gilbert Robert. Tagger Overview. Rapport technique, ISSCO, Genève, Suisse, 1995.

[Aga95] Rajeev Agarwal. *Semantic Feature Extraction from Technical Texts with Limited Human Intervention.* PhD thesis, Mississippi State University, États-Unis, 1995.

[Ant90] Evan L. Antworth. PC-KIMMO: A Two-Level Processor for Morphological Analysis. Technical Report 16, Summer Institute of Linguistics, Dallas, États-Unis, 1990.

[Arm96] Susan Armstrong. Multext: Multilingual Text Tools and Corpora. In H. Feldweg and W. Hinrichs, editors, *Lexikon und Text*, pages 107–119. Tübingen: Niemeyer, 1996.

[Ass98] Houssem Assadi. *Construction d'ontologies à partit de textes techniques - Applications aux systèmes documentaires.* Thèse de doctorat, Université de Paris 6, France, 1998.

[BB01] Pierrette Bouillon and Federica Busa, editors. *Generativity in the Lexicon.* CUP:Cambridge, 2001.

[BBRR00] Pierrette Bouillon, Robert H. Baud, Gilbert Robert, and Patrick Ruch. Indexing by Statistical Tagging. In *5^es Journées d'analyse statistique de données textuelles, JADT'2000*, Lausanne, Suisse, 2000.

[BC97] Ted Briscoe and John Carroll. Automatic Extraction of Subcategorisation from Corpora. In *5th ACL Conference on Applied Natural Language Processing, ANLP97*, Washington, États-Unis, 1997.

[BC99] Didier Bourigault and Anne Condamines. Alternance nom/verbe : explorations en corpus spécialisés. In B. Victorri and J. François, editors, *Sémantique du lexique verbal*, Cahiers de l'Elsap, Caen, France, 1999.

[BCFS01] Pierrette Bouillon, Vincent Claveau, Cécile Fabre, and Pascale Sébillot. Using Part-of-Speech and Semantic Tagging for the Corpus-Based Learning of Qualia Structure Elements. In P. Bouillon and

K. Kanzaki, editors, *1st International Workshop on Generative Approaches to the Lexicon, GL'2001*, Genève, Suisse, 2001.

[BCL01] Federica Busa, Nicoletta Calzolari, and Alessandro Lenci. Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP. In F. Busa and P. Bouillon, editors, *Generativity in the Lexicon*, chapter 21, pages 387–405. CUP:Cambridge, 2001.

[Beu98] Pierre Beust. *Contribution à un modèle interactionniste du sens*. Thèse de doctorat, Université de Caen, France, 1998.

[BFSJ00] Pierrette Bouillon, Cécile Fabre, Pascale Sébillot, and Laurence Jacqmin. Apprentissage de ressources lexicales pour l'extension de requêtes. *TAL (Traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, 41(2):367–393, 2000.

[BHNZ97] Jacques Bouaud, Benoît Habert, Adeline Nazarenko, and Pierre Zweigenbaum. Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation avec deux modélisations conceptuelles. In *Ingénierie de la Connaissance*, Roscoff, France, 1997.

[BJL01] Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors. *Recent Advances in Computational Terminology*. John Benjamins, Cambridge MA, 2001.

[BK01] P. Bouillon and K. Kanzaki, editors. *Proceedings of the 1st International Workshop on Generative Approaches to the Lexicon, GL'2001*, Genève, Suisse, 2001.

[BM98] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-training. In *Workshop on Computational Learning Theory, COLT*, Carnegie Mellon University, Pittsburgh, États-Unis, 1998. Morgan Kaufmann.

[Bou02] Didier Bourigault. Analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Traitement automatique des langues naturelles, TALN'02*, Nancy, France, 2002.

[BS99] Liviu Badea and Monica Stanciu. Refinement Operator can be (Weakly) Perfect. In *9th International Conference on Inductive Logic Programming, ILP-99*, Bled, Slovénie, 1999.

[Bui97] Paul Buitelaar. A Lexicon for Underspecified Semantic Tagging. In *ANLP'97 Workshop on Tagging text with Lexical Semantics*, Washington, États-Unis, 1997.

[Bun88] Wray Lindsay Buntine. Generalized Subsumption and its Application to Induction and Redundancy. *Artificial Intelligence*, 36:375–399, 1988.

[Bur90] Gavin Burnage. *CELEX: A Guide for Users*. Center for Lexical Information, University of Nijmegen, Pays-Bas, 1990. http://www.kun.nl/celex/.

[Cat02] Emmanuelle Catz. Apprentissage automatique de catégories de mots par co-training. Rapport de DEA, IFSIC, Université de Rennes 1, France, 2002.

[CG91] Kenneth W. Church and William A. Gale. Concordances for Parallel Texts. In *7th Annual Conference of the UW Centre for the New OED and Text Research*, University of Waterloo, Ontario, Canada, 1991.

[CHU00] Special Issue on Senseval. *Computers and the Humanities*, 34(1/2), 2000.

[Cru86] D. Alan Cruse. *Lexical Semantics*. Cambridge Textbooks in Linguistics, 1986.

[CS89] Blandine Courtois and Max Silberztein. Les dictionnaires électroniques DELAS et DELAC. In *Colloque sur les Langues Romanes*, Université Laval, Québec, Canada, 1989. ASTRIL-LADL.

[CSBF01] Vincent Claveau, Pascale Sébillot, Pierrette Bouillon, and Cécile Fabre. Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ? *TAL (Traitement automatique des langues), numéro spécial Lexiques sémantiques dans les applications du TAL*, 42(3):729–753, 2001.

[CSFB02] Vincent Claveau, Pascale Sébillot, Cécile Fabre, and Pierrette Bouillon. Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming. *JMLR (Journal of machine learning research), special issue on inductive logic programming, à paraître*, 2002.

[Dai94] Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*. Thèse de doctorat, Université Paris VII, France, 1994.

[Dai00] Béatrice Daille. Morphological Rule Induction for Terminology Acquisition. In *18th International Conference on Computational Linguistics, COLING 2000*, Saarbrücken, Allemagne, 2000.

[Dai02] Béatrice Daille. Découvertes linguistiques en corpus. Habilitation à diriger des recherches, Université de Nantes, France, 2002.

[DFS02] Béatrice Daille, Cécile Fabre, and Pascale Sébillot. Applications of Computational Morphology. In P. Boucher, editor, *Many Morphologies*, pages 210–234. Cascadilla Press, Somerville, 2002.

[dG00] Édith Galy. Repérer en corpus les associations sémantiques privilégiées entre le nom et le verbe : le cas de la fonction dénotée par le nom. Mémoire de Maîtrise, Université de Toulouse - Le Mirail, France, 2000.

[DRF89] Fathi Debili, Pierre Radasoa, and Christian Fluhr. About Reformulation in Full-Text IRS. *Information Processing and Management*, 25(6):647–657, 1989.

[Dun93] Ted E. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[ELMS96] Floriana Esposito, Angela Laterza, Donato Malerba, and Giovanni Semeraro. Refinement of Datalog Programs. In *MLnet Familiarization Workshop on Data Mining with Inductive Logic Programming*, Bari, Italie, 1996.

[Fab96] Cécile Fabre. *Interprétation automatique des séquences binominales en anglais et en français. Application à la recherche d'informations.* Thèse de doctorat, Université de Rennes 1, France, 1996.

[Fag87] Joel L. Fagan. *Experiments in Automatic Phrase Indexing for Document Retrieval: a Comparison of Syntactic and Non-Syntactic Methods.* PhD thesis, Cornell University, Ithaca, États-Unis, 1987.

[Fel98] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[FG01] Olivier Ferret and Brigitte Grau. Utiliser des corpus pour amorcer une analyse thématique. *TAL (Traitement automatique des langues), numéro spécial Linguistique de corpus*, 42(2):517–545, 2001.

[FHL97] Cécile Fabre, Benoît Habert, and Dominique Labbé. La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*, 13:15–30, 1997.

[FJ00] Cécile Fabre and Christian Jacquemin. Boosting Variant Recognition with light Semantics. In *18th International Conference on Computational Linguistics, COLING 2000*, Saarbrücken, Allemagne, 2000.

[FN99] David Faure and Claire Nédellec. Knowledge Acquisition of Predicate Argument Structures from Technical Texts using Machine Learning: the System ASIUM. In Dieter Fensel Rudi Studer, editor, *11th European Workshop EKAW'99*, Dagstuhl, Allemagne, 1999. Springer-Verlag.

[Fol02] Helka Folch. *Articuler les classifications sémantiques induites d'un domaine.* Thèse de doctorat, Université de Paris-Sud, France, 2002.

[FS99] Cécile Fabre and Pascale Sébillot. Semantic Interpretation of Binominal Sequences and Information Retrieval. In *International ICSC Congress on Computational Intelligence: Methods and Applications, CIMA'99, Symposium on Advances in Intelligent Data Analysis, AIDA'99*, Rochester, États-Unis, 1999.

[GGHR00] Éric Gaussier, Gregory Grefenstette, David Hull, and Claude Roux. Recherche d'information en français et traitement automatique des langues. *TAL (Traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, 41(2):473–493, 2000.

[GGS97] Éric Gaussier, Gregory Grefenstette, and Maximilian B. Schulze. Traitement du Langage Naturel et Recherche d'Informations : quelques expériences sur le français. In *1es Journées scientifiques et techniques du réseau FRANCIL de l'AUPELF-UREF, JST'97*, Avignon, France, 1997.

[GJ93] Danièle Godard and Jacques Jayez. Towards a Proper Treatment of Coercion Phenomena. In *European Chapter of the Association fo Computational Linguistics, EACL*, Utrecht, Pays-Bas, 1993.

[GLO93] Michel Gilloux, Edmond Lassalle, and Jean-Michel Ombrouck. Interrogation en langage naturel du Minitel guide des services. *Écho des recherches*, 146:1–20, 1993.

[Gre94a] Gregory Grefenstette. Corpus-Derived First, Second and Third-Order Word Affinities. In *EURALEX'94*, Amsterdam, Pays-Bas, 1994.

[Gre94b] Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Dordrecht: Kluwer Academic Publishers, 1994.

[Gre97] Gregory Grefenstette. SQLET: Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text. In McGill-University, editor, *Recherche d'Informations Assistée par Ordinateur, RIAO'97*, Montréal, Québec, Canada, 1997.

[GT95] Gregory Grefenstette and Simone Teufel. Corpus-Based Method for Automatic Identification of Support Verbs for Nominalizations. In 7^{th} *Conference of European Chapter of the Association for Computational Linguistics*, Dublin, Irlande, 1995.

[GZ99] Natalia Grabar and Pierre Zweigenbaum. Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Traitement automatique des langues naturelles, TALN'99*, Cargèse, France, 1999.

[Har91] Donna Harman. How Effective is Suffixing? *JASIS: Journal of the American Society for Information Science*, 42:7–15, 1991.

[Hea92] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *15th International Conference on Computational Linguistics, COLING-92*, Nantes, France, 1992.

[Hea94] Marti A. Hearst. Multi-Paragraph Segmentation of Expository Texts. In *32th Annual Meeting of the Association for Computational Linguistics, ACL'94*, Las Cruces, États-Unis, 1994.

[Hea98] Marti A. Hearst. Automatic Discovery of WordNet Relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 5, pages 131–151. MIT Press, Cambridge MA, 1998.

[HGR+89] Zellig Harris, Michael Gottfried, Thomas Ryckman, Paul Mattick(Jr), Anne Daladier, Tzvee N. Harris, and Suzanna Harris. The Form of Information in Science, Analysis of Immunology Sublanguage. *Boston Studies in the Philosophy of Science*, 104, 1989.

[HNS97] Benoît Habert, Adeline Nazarenko, and André Salem. *Les linguistiques de corpus.* Armand Collin/Masson, Paris, 1997.

[IV94] Nancy Ide and Jean Véronis. MULTEXT (Multilingual Tools and Corpora). In *15th International Conference on Computational Linguistics, COLING-94*, Kyoto, Japon, 1994.

[IV98] Nancy Ide and Jean Véronis. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24(1):1–40, 1998.

[Jac96] Christian Jacquemin. A Symbolic and Surgical Acquisition of Terms through Variation. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438. Springer, Heidelberg, 1996.

[Jac97] Christian Jacquemin. Guessing Morphology from Terms and Corpora. In *20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'97*, Philadelphia, États-Unis, 1997.

[Jac00] Christian Jacquemin. Présentation. *TAL (Traitement automatique des langues), numéro spécial traitement automatique des langues pour la recherche d'information*, 41(2):327–332, 2000.

[Jac01] Christian Jacquemin. *Spotting and Discovering Terms through NLP.* MIT Press, Cambridge MA, 2001.

[JKT97] Christian Jacquemin, Judith L. Klavans, and Evelyne Tzoukermann. Expansion of Multi-Word Terms for Indexing and Retrieval Using Morphology and Syntax. In *35th Annual Meeting of the Association for Computational Linguistics, ACL'97*, Madrid, Espagne, 1997.

[JR93] Paul S. Jacobs and Lisa F. Rau. Innovations in Text Interpretation. In Fernando C.N. Pereira and Barbara J. Grosz, editors, *Natural Language Processing*, pages 143–191. MIT/Elsevier, 1993.

[JT99] Christian Jacquemin and Evelyne Tzoukermann. NLP for Term Variant Extraction: Synergy of Morphology, Lexicon, and Syntax. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25–74. Kluwer, Boston, MA, 1999.

[Kil01] Adam Kilgarriff. How Much of the Time does the Generative Lexicon Account for Novel Word Uses? In P. Bouillon and K. Kanzaki, editors, *1st International Workshop on Generative Approaches to the Lexicon, GL'2001*, Genève, Suisse, 2001.

[KK98] Judith Klavans and Min-Yen Kan. Role of Verbs in Document Analysis. In *17th International Conference on Computational Linguistics and Association for Computational Linguistics, COLING-ACL*, Montréal, Québec, Canada, 1998.

[Koh95] Ron Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *14th International Joint Conference on Artificial Intelligence, IJCAI 95*, Montréal, Québec, Canada, 1995.

[Kro97] Robert Krovetz. Homonymy and Polysemy in Information Retrieval. In *35th Annual Meeting of the Association for Computational Linguistics, ACL'97*, Madrid, Espagne, 1997.

[Ler91] Israël-César Lerman. Foundations in the Likelihood Linkage Analysis Classification Method. *Applied Stochastic Models and Data Analysis*, 7:69–76, 1991.

[Lov68] Julie B. Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[LP95] Diane J. Litman and Rebecca J. Passonneau. Combining Multiple Knowledge Sources for Discourse Segmentation. In *33th annual meeting of the Association for Computational Linguistics, ACL'95*, Montréal, Québec, Canada, 1995.

[LPTW81] Martin Lennon, David S. Pierce, Brian D. Tarry, and Peter Willet. An Evaluation of some Conflation Algorithms for Information Retrieval. *Journal of Information Science*, 3:177–183, 1981.

[MDR94] Stephen Muggleton and Luc De Raedt. Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19-20:629–679, 1994.

[Moo99] Raymond J. Mooney. Learning for Semantic Interpretation: Scaling Up Without Dumbing Down. In *Learning Language in Logic Workshop, LLL'99*, Bled, Slovénie, 1999.

[Mor97] Emmanuel Morin. Extraction de liens sémantiques entre termes dans des corpus de textes techniques : application à l'hyponymie. In *Traitement Automatique des Langues Naturelles, TALN'97*, Grenoble, France, 1997.

[Mor99] Emmanuel Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* Thèse de doctorat, Université de Nantes, France, 1999.

[Mug95] Stephen Muggleton. Inverse Entailment and Progol. *New Generation Computing*, 13(3-4):245–286, 1995.

[Nam00] Fiammetta Namer. Flemm : un analyseur flexionnel du français à base de règles. *TAL (Traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, 41(2):523–547, 2000.

[Néd99] Claire Nédellec. Corpus-Based Learning of Semantic Relations by the ILP System, Asium. In *Learning Language in Logic Workshop, LLL'99*, Bled, Slovénie, 1999.

[NRA+96] Claire Nédellec, Céline Rouveirol, Hilde Adé, Francesco Bergadano, and Birgit Tausend. Declarative Bias in Inductive Logic Programming. In Luc De Raedt, editor, *Advances in Inductive Logic Programming*, pages 82–103. IOS Press, 1996.

[NZHB01] Adeline Nazarenko, Pierre Zweigenbaum, Benoît Habert, and Jacques Bouaud. Corpus-Based Extension of a Terminological Semantic Lexicon. In Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, editors, *Recent Advances in Computational Terminology*, chapter 16, pages 327–351. John Benjamins Publishing Company, 2001.

[PAB93] James Pustejovsky, Peter Anick, and Sabine Bergler. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics*, 19(2), 1993.

[PBV+97] James Pustejovsky, Branimir Boguraev, Marc Verhagen, Paul Buitelaar, and Michael Johnston. Semantic Indexing and Typed Hyperlin-

king. In *AAAI Spring 1997 Workshop on Natural Language Processing for the World Wide Web*, Stanford, États-Unis, 1997.

[Pin99] Bénédicte Pincemin. Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative. In *Atelier thématique "Corpus et TAL : pour une réflexion méthodologique", Traitement automatique des langues naturelles, TALN'99*, Cargèse, France, 1999.

[PLL92] Philippe Peter, Henri Leredde, and Israël-César Lerman. *Notice du programme CHAVL*, 1992.

[Plo70] Gordon D. Plotkin. A Note on Inductive Generalization. *Machine Intelligence*, 5:153–163, 1970.

[Por80] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14:130–137, 1980.

[PR94] Dominique Petitpierre and Graham Russell. Mmorph - the Multext Morphology Program. Rapport technique, ISSCO, Genève, Suisse, 1994.

[PS97] Ronan Pichon and Pascale Sébillot. Acquisition automatique d'informations lexicales à partir de corpus : un bilan. Rapport de recherche n^{0}3321, INRIA, Rennes, France, 1997.

[PS99] Ronan Pichon and Pascale Sébillot. Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. In *Traitement automatique des langues naturelles, TALN'99*, Cargèse, France, 1999.

[PS00] Ronan Pichon and Pascale Sébillot. From Corpus to Lexicon: from Contexts to Semantic Features. In Barbara Lewandowska-Tomaszczyk and Patrick James Melia, editors, *PALC'99: Practical Applications in Language Corpora*, volume 1 of *Lodz studies in Language*. Peter Lang, 2000.

[Pus95] James Pustejovsky. *The Generative Lexicon*. Cambridge: MIT Press, 1995.

[PW95] Marie-Paule Péry-Woodley. Quels corpus pour quels traitements automatiques? *TAL (Traitement automatique des langues), numéro spécial Traitement probabilistes et corpus*, 36(1-2):213–232, 1995.

[Qui90] John R. Quinlan. Learning Logical Definitions from Relations. *Machine Learning*, 5:239–266, 1990.

[Ras95] François Rastier. La sémantique des thèmes ou le voyage sentimental. In *L'analyse thématique des données textuelles*, pages 223–249. Didier, Paris, 1995.

[Ras96] François Rastier. *Sémantique Interprétative*. Presses Universitaires de France, seconde édition, 1996.

[RBB+99] Patrick Ruch, Pierrette Bouillon, Robert H. Baud, Anne-Marie Rassinoux, and Jean-Raoul Scherrer. MEDTAG: Tag-like Semantics for Medical Document Indexing. In *American Medical Informatics Association, AMIA99 Annual Symposium*, Washington, États-Unis, 1999.

[RBC00] Martin Rajman, Romaric Besançon, and Jean-Cédric Chappelier. Le modèle DSIR : une approche à base de sémantique distributionnelle pour la recherche documentaire. *TAL (Traitement automatique des langues), numéro spécial Traitement automatique des langues pour la recherche d'information*, 41(2):549–578, 2000.

[RCA94] François Rastier, Marc Cavazza, and Anne Abeillé. *Sémantique pour l'analyse : de la linguistique à l'informatique*. Masson, 1994.

[Res93] Philip S. Resnik. *Selection and Information: a Class-Based Approach to Lexical Relationships*. PhD thesis, University of Pennsylvania, États-Unis, 1993.

[Res95a] Philip Resnik. Disambiguating Noun Groupings with Respect to WordNet Senses. In 3^{rd} *Workshop on Very Large Corpora, Association for Computational Linguistics*, Cambridge, États-Unis, 1995.

[Res95b] Philip Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *14th International Joint Conference on Artificial Intelligence, IJCAI'95*, Montréal, Québec, Canada, 1995.

[Ros01] Mathias Rossignol. Acquisition sur corpus d'informations lexicales basées sur la sémantique différentielle. Rapport de DEA, IFSIC, Université de Rennes 1, France, 2001.

[RS02] Mathias Rossignol and Pascale Sébillot. Automatic Generation of Sets of Keywords for Theme Characterization and Detection. In A. Morin and P. Sébillot, editors, 6^{es} *Journées internationales d'analyse statistique des données textuelles, JADT'2002*, Saint-Malo, France, 2002.

[Sal89] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. New York: Addison-Wesley, 1989.

[SBC+00] Pascale Sébillot, Pierrette Bouillon, Vincent Claveau, Cécile Fabre, Laurence Jacqmin, and Jacques Nicolas. Apprentissage en corpus de couples nom-verbe pour la construction d'un lexique génératif. In 5^{es} *Journées d'analyse statistique de données textuelles, JADT'2000*, Lausanne, Suisse, 2000.

[SBF00] Pascale Sébillot, Pierrette Bouillon, and Cécile Fabre. Inductive Logic Programming for Corpus-Based Acquisition of Semantic Lexicons. In *Learning Language in Logic, LLL-2000*, Lisbonne, Portugal, 2000.

[Sha81] Ehud Y. Shapiro. Inductive inference of theories from facts. Rapport de recherche 624, Department of Computer Science, Yale University, New Haven, États-Unis, 1981.

[SJ99] Karen Spärck Jones. What is the Role of NLP in Text Retrieval? In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 1–24. Kluwer Academic Publishers, 1999.

[SJT84] Karen Spärck Jones and John I. Tait. Automatic Search Term Variant Generation. *Journal of Documentation*, 40(1):50–66, 1984.

[SLWPC99] Tomek Strzalkowski, Fang Lin, Jin Wang, and Jose Perez-Carballo. Evaluating Natural Language Processing Techniques in Information

Retrieval. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 113–145. Kluwer Academic Publishers, 1999.

[Sme99] Alan F. Smeaton. Using NLP or NLP Resources for Information Retrieval Tasks. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 99–111. Kluwer Academic Publishers, 1999.

[Str99] Tomek Strzalkowski. Preface. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages xiii–xxii. Kluwer Academic Publishers, 1999.

[Tan97] Ludovic Tanguy. *Traitement Automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative.* Thèse de doctorat, École nationale supérieure des télécommunications de Bretagne - Université de Rennes 1, France, 1997.

[Tar00] Olivier Tardif. Acquisition automatique de lexiques sémantiques basés sur la sémantique différentielle. Rapport de DEA, IST Génie linguistique, Université de Marne-La Vallée, France, 2000.

[TR97a] Fabien Torre and Céline Rouveirol. Natural Ideal Operators in Inductive Logic Programming. In M. van Someren and Widmer G., editors, *9th European Conference on Machine Learning, ECML'97, LNAI 1224*, volume 1224, Prague, République Tchèque, 1997. Springer-Verlag.

[TR97b] Fabien Torre and Céline Rouveirol. Opérateurs naturels en programmation logique inductive. In *12es Journées françaises d'apprentissage, JFA'97*, Roscoff, France, 1997.

[TR97c] Fabien Torre and Céline Rouveirol. Private Properties and Natural Relations in Inductive Logic Programming. Rapport technique 1118, Laboratoire de Recherche en Informatique d'Orsay, France, 1997.

[Van00] Laurence Vandenbroucke. Indexation automatique par couples nom-verbe pertinents. Rapport de DES en information et documentation, Faculté de Philosophie et Lettres, Université Libre de Bruxelles, Belgique, 2000.

[vdL95] Patrick R.J. van der Laag. *An Analysis of Refinement Operators in Inductive Logic Programming.* PhD thesis, Erasmus Universiteit, Rotterdam, Pays-Bas, 1995.

[VFP91] Paola Velardi, Michela Fasolo, and Maria Teresa Pazienza. How to Encode Semantic Knowledge: a Method for Meaning Representation and Computer-Aided Acquisition. *Computational Linguistics*, 17(2):153–170, 1991.

[Voo94] Ellen M. Voorhees. Query Expansion using Lexical-Semantic Relations. In *ACM SIGIR'94*, Dublin, Irlande, 1994.

[Voo98] Ellen M. Voorhees. Using WordNet for Text Retrieval. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, chapter 12, pages 285–303. MIT Press, Cambridge MA, 1998.

[Vos98] Piek Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, Dordrecht, 1998.

[WRS96] Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*. Lecture Notes in Computer Science, vol. 1040, Springer-Verlag, 1996.

[WS96] Yorick Wilks and Mark Stevenson. The Grammar of Sense: is Word-Sense Tagging much more than Part-of-Speech Tagging? Rapport technique, University of Sheffield, Grande-Bretagne, 1996.

[WSG96] Yorick A. Wilks, Brian M. Slator, and Louise Guthrie. *Electric Words: Dictionaries, Computers, and Meanings*. Bradford, 1996.

[XC98] Jinxi Xu and Bruce W. Croft. Corpus-Based Stemming using Cooccurrence of Word Variants. *ACM Transactions on Information Systems*, 16(1):61–81, 1998.

[Yar95] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, États-Unis, 1995.